Modelling and Measuring Data Quality for Quality-Awareness in Data Mining

Laure Berti-Équille

IRISA, Campus Universitaire de Beaulieu, Rennes, France
Laure.Berti-Equille@irisa.fr

Abstract. This paper presents an overview of data quality management and data cleaning techniques that can be advantageously employed for improving the quality awareness of knowledge discovery processes. It proposes a framework for data quality enhancement and awareness before warehousing and mining data. Finally, a cost model that predicts the cost of low-quality data on the quality of discovered association rules is proposed.

Keywords: Data Quality Management, Data Cleaning, Data Quality Metadata, Association Rules, Cost Model

1. Introduction

The quality of data mining results and the validity of results interpretations essentially rely on the data preparation process and on the quality of the analyzed datasets. Indeed, data mining processes and applications require various forms of data preparation, correction and consolidation, combining complex data transformation operations and cleaning techniques. This is because the data input to the mining 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 and analyze the data for discovering added-value knowledge and decision making.

Data quality is a multidimensional, complex and morphing concept [17]. In the last decade, there has been a significant amount of work in the area of information and data quality management initiated by several research communities (database, statistics, workflow management, knowledge engineering), ranging from techniques that assess information quality to build large-scale data integration systems over heterogeneous data sources with different degrees of quality and trust. Many data quality definitions, metrics, models and methodologies have been proposed by academics and practitioners with the aim to tackle the main classes of data quality problems:

- **Duplicate detection** and **record matching** known under various names: record linkage [27], merge/purge problem [33], object matching [15,88], duplicate elimination [1,47,51], citation matching [7,48], identity uncertainty [59], entity identification [43], entity resolution [6], or approximate string join [31],
- **Instance conflict resolution** [26] using data source selection [19,53] or data cleaning techniques [66,89],
- **Missing values** [44] and **incomplete data** [72],

- **Staleness of data** [8,16].

In error-free data warehouses or database-backed information systems with perfectly clean data, knowledge discovery techniques (such as clustering, mining association rules or visualization) can be relevantly used as decision making processes to automatically derive new knowledge patterns and new concepts from data. Unfortunately, most of the time, this data is neither rigorously chosen from the various heterogeneous sources with different degrees of quality and trust, nor carefully controlled for quality. Deficiencies in data quality still are a burning issue in many application areas, and become acute for practical applications of knowledge discovery and data mining techniques [60]. An example using association rules discovery is used to illustrate this basic idea. This will be completed by other application examples in Section 2. Among traditional descriptive data mining techniques, association rules discovery identifies intra-transaction patterns in a database and describes how much the presence of a set of attributes in a database's record (*i.e.*, a transaction) implicates the presence of other distinct set of attributes in the same record (respectively the same transaction). The quality of association rules is commonly evaluated by their support and confidence. The support of a rule measures the occurrence frequency of the pattern in the rule while the confidence is the measure of the strength of implication. The problem of mining association rules is to generate all association rules that have support and confidence greater than the user-specified minimum support and confidence thresholds. Besides support and confidence, other measures for knowledge quality evaluation (called interestingness measures) have been proposed in the literature with the purpose of supplying indicating alternatives to the user in the understanding and use of the new discovered knowledge [41,77]. But, to illustrate the impact of low-quality data over discovered association rule quality, one might legitimately wonder whether a so-called "interesting" rule with the Left-Hand Side and Right-Hand Side of the rule noted *LHS* □ *RHS* is meaningful when 30 % of the *LHS* data are not up-to-date anymore, 20% of the *RHS* data are not accurate, and 15% of the *LHS* data come from a data source that is well-known for its bad reputation and lack of credibility.

The main contribution of this chapter is threefold: first, an exhaustive overview of data quality management is given and can be advantageously employed for improving the data quality awareness of knowledge discovery and data mining techniques; secondly, a framework for data quality enhancement and awareness for the KDD process is proposed; finally, a cost model is described to predict the cost of low-quality data over the quality of discovered association rules.

The rest of the chapter is organized as follows. Section 2 discusses motivations for data quality awareness and management in three mining application contexts. Section 3 gives an exhaustive overview on data quality characterization, management and cleaning techniques. In Section 4, a quality-aware KDD framework is presented and also a decision model for estimating the cost of low-quality data on mining association rules. Section 5 provides concluding remarks and guidelines for future extensions of this work.

2. Motivation

Three mining application areas are presented to show the importance of data quality awareness: *i.e.*, data accuracy, precision, completeness, currency, trustworthiness, non-duplication, and data source reputation.

Application 1: In life sciences, researchers extensively collaborate with each other, sharing biomedical and genomic data and their experimental results. This necessitates dynamically integrating different databases or warehousing them into a single repository [32]. Overlapping data sources may be maintained in a controlled way, such as replication of data on different sites for load balancing or for security reasons. But uncontrolled overlaps are very frequent cases. Moreover, scientists need to know how reliable the data is if they are to base their research on it, because pursuing incorrect theories and experiments costs time and money. The current solution to ensure data quality is verification by human experts. The two main drawbacks are: *i)* data sources are autonomous and as a result, sources may provide excellent reliability in one specific area, or on some data subsets, and *ii)* curation is a manual process of data accreditation by specialists that slows the incorporation of data and that is not free from conflicts of interest. Biological databank providers will not directly support data quality evaluations to the same degree since there is no motivation for them to and there are currently no standards for evaluating and comparing biomedical data quality. Mining for patterns in contradictory biomedical data has been proposed (e.g., [52]), but more automatic, impartial, and independent data quality evaluation techniques are needed for all types of data before any biomedical mining applications. Moreover, chips and micro-array data are now becoming standard tools for the high-throughput analysis of gene expression. The power of micro-array analysis and array-based gene expression image mining lies in the ability to compare large sets of genes in different tissues or conditions to identify pathways and regulatory networks. As micro-array laboratories are scaling up their facilities, manual assessment of chip images becomes cumbersome and prone to subjective criteria. Several systems exist to standardize data quality control, but many others need to be developed for the highly automated assessment and optimization of experimental micro-array data quality and, therefore for monitoring a variety of quality parameters and offering means to minimize noise, remove systematic measurement errors before the bio-mining processes.

Application 2: In business or technological intelligence gathering efforts, data is often collected from many heterogeneous data sources, such as technical reports, human assets, transcripts, commercial documents, competitive studies, knowledge-sharing web sites, newsgroups, etc. It is obvious that each of these data sources have different degrees of trust and quality that can be advantageously (even maliciously) used (e.g., in viral marketing). With the variety of data sources to consider and incorporate, it is currently time-consuming to sift through each of these data sources to determine which are the most accurate. To make the correct decisions based on the intelligence available in a timely manner, automatic means are needed to determine accurate data sources and to be able to detect malicious or compromised data sources to prevent them from influencing the decision making processes. Mining techniques should be aware of completeness, accuracy, trustworthiness and inter-dependency of data sources for ensuring critical data and decisions.

Application 3: Multimedia content (combining audio, video, image, text) is prevailing for all data exchanges, transferred through various heterogeneous networks (IP, DVB, DVB over IP, IP over DVB) to several kinds of terminal devices (TV set, set top box, mobile phone, smart phone, PDA, PC). Ensuring end-to-end quality of service along the whole audio-visual delivery chain and integrity of multimedia content through adaptation to network and terminal characteristics is the commonly shared goal of all the content, service and network providers today in the market. These characteristics are mainly captured by sensors and may be contradictory or erroneous. Human involvement in the raw data capturing process can also introduce noise. On the other side of the multimedia delivery chain, multimedia mining techniques try to decipher the multimedia data deluge trying to bridge the "semantic" gap. High-dimension reduction techniques, such as pre-clustering multimedia databases rely on media description accuracy and freshness because these descriptive features (usually sensitive to the dataset size) are pre-computed only once over very large multimedia databases. In most cases, they are not synchronously updated and re-computed each time the database is modified. Clustering and association rules discovery on multimedia databases [22,57,61,91] generally require the manual creation of a thesaurus of labels (e.g., "animal", "plant", etc.) for semantically describing groups of images or videos and then, inferring the mapping between the keywords labels (at the semantic level) and the clusters of low-level descriptive features (at the pixel level). Here again, human subjectivity may lead to misinterpretation or bad precision and recall for the content-based information retrieval scenarios. Low-quality of multimedia mining results is therefore due to the lack of completeness and precision of the chosen key-words for characterizing the richness of multimedia contents.

As briefly shown above, these three mining application contexts are conditioned to various data quality problems occurring along the entire data processing continuum [17]. Data preparation plays a major role with several necessary operations such as cleaning data, normalizing, handling noisy, uncertain or untrustworthy information, handling missing values, transforming and coding data in such a way that data become suitable for the data mining process. As shown in Table 1, synthesizing Dasu and Johnson's vision [17], several academic or commercial solutions have been proposed to tackle more or less pragmatically the continuous problems of data quality at each step of the data processing.

3. An Overview of Data Quality Management

Maintaining a certain level of quality of data is challenging and cannot be limited to one-shot approaches addressing simpler, more abstract versions of the problems of dirty or low-quality data [14,17,42].

PROCESSING STEP	DATA QUALITY PROBLEMS	POTENTIAL SOLUTIONS AND REFERENCES
Data Creation, Capture, Gathering or Import	Manual entry Complex Data Type (image, audio, video, text) No standardized format or data schema Duplicates Approximations, surrogates Hardware or software constraints or limitations Measurement / Sensor Errors Automatic massive data import	Preemptive Approaches: – Workflow Management Methodologies [25,68,86,87] – Architectures for Data Quality Management [4,20,35,46,58,76] – Data audits, data stewardship [12] Retrospective and Corrective Approaches: – *Data Diagnosis* : error and outliers detection [9,40] – *Data Cleaning:* record linkage [27,89], merge/purge problem [33], object matching [15,88], duplicate elimination [1,47,51], name disambiguation, citation matching [7,48], identity uncertainty [59], entity identification [43], entity resolution [6], or approximate string join [31], address and string matching [55]
Data Delivery	Information destruction or mutilation by inappropriate pre-processing Data Loss: buffer overflows, transmission problems No Checks	- Use checksum or data mining techniques to check correctness of data transmissions - Monitor data transmission, data integrity, data format - Data Quality control [42] - Data editing - Data publishing - Data aggregation and data squashing [23]
Data Storage	Metadata paucity and staleness Inappropriate data models and schemas *Ad hoc* Modifications Hardware of Software constraints or limitations	- Metadata Management [17,69] - Plan ahead and customize for domain: Data Profiling, Data browsing to monitor [18,92]
Data Integration	Multiple heterogeneous data sources Time synchronization Atypical Data Legacy systems Sociological factors *Ad hoc* Joints Random Matching Heuristics	- Mandate accurate timestamps, data lineage [16] - Commercial tools for data migration - Data scrubbing, profiling [13] - Academic tools and language extensions for data cleaning (ETL): Potter's Wheel [67], Ajax [29], Bellman [18], Arktos [81] and quality-driven query processing: HIQIQ [53,54] - Academic tools for approximate join matching
Data Retrieval	Human errors Computational constraints Software constraints or limitations, incompatibility Recall / Precision significance	- Feedback loop
Statistical Analysis & Data Mining	Issues of scale, performance and confidence guarantees Belief in black boxes and dart boards Attachment to a family of models Insufficient domain expertise Lack of familiarity with the data	- Data Preparation for mining [60,64] - *Exploratory Data Mining* œ(EDM) [17] - Greater accountability from analysts - Continuous, ongoing analysis rather than one-shot solutions - Sampling vs. full analysis - Feedback loops

Table 1 – *Data Quality: Problems and Current Solutions for Data Quality Management*

Solving these problems requires highly domain- and context-dependent information and human expertise. Classically, the database literature refers to data quality management as ensuring: *i)* syntactic correctness (e.g., constraints enforcement that prevent "garbage data" from being entered into the database) and *ii)* semantic correctness (*i.e.*, data in the database that truthfully reflects the real world situation). This traditional approach of data quality management has lead to techniques such as integrity constraints, concurrency control and schema integration for distributed and heterogeneous systems. A broader vision of data quality management is presented in this chapter (but still with a database orientation).

In the last decade, literature on data and information quality across different research communities (including databases, statistics, workflow management and knowledge engineering) have proposed a plethora of:

- **Data quality dimensions** with various definitions depending on authors and application contexts [5, 20, 28,85],
- **Data quality dimension classifications** that are depending on the audience type: practical and technical [68]or more general [39] or depending on the architecture type: [54] for integrated information systems and [37] for data warehouse systems or [70] for cooperative information systems (CIS),

- **Data quality metrics** [17,34,54,62,63,90],
- **Conceptual models** [50,62,63,71,83],
- **Frameworks** and **methodologies** to improve or assess data quality in databases [2,35,68,85,86], in (cooperative) information systems [3,4,87] or in data warehouse systems [25,36,80,82].

To give a detailed overview, the rest of this section will present the different paradigms for data quality characterization, modeling, measurement and management methodologies.

3.1. Data Quality Characterization, Measures, Models and Management Methodologies

Data Quality Dimensions

Since 1980 with [10], more than 200 dimensions have been collected to characterize data quality in the literature [5,35,68,84]. The most frequently mentioned data quality dimensions in the literature are accuracy, completeness, timeliness and consistency with various definitions:

- Accuracy is the extent to which collected data are free of error measurements [45] or can be measure by the quotient of the number of correct values in a source and the number of the overall number of values [54],
- Completeness is the percentage of the real-world information entered in the sources and/or the data warehouse [36] or measured by the quotient of the number of non-null values in a source and the size of the universal relation [54],
- Timeliness is the extent to which data are sufficiently up-to-date for a task [45]; others definitions of freshness, currency, volatility are reported in [8],
- Consistency is the coherence of the same data represented in multiple copies or different data with respect of integrity constraints and rules [5].

Figure 1 presents a classification of data quality dimensions divided into the four following categories:

- Quality of the management of data by the system based on the satisfaction of technical and physical constraints (e.g., accessibility, ease of maintenance, reliability, etc.),
- Quality of the representation of data in the system based on the satisfaction of conceptual constraints on modeling and information presentation (e.g., conformance to schema, appropriate presentation, clarity, etc.),
- Intrinsic data quality dimensions (e.g., accuracy, uniqueness, consistency, etc.),
- Relative data quality dimensions with dependence on the user (e.g., user preferences), or on the application (e.g., criticality, conformance to business rules, etc.), or on time (e.g., variability, volatility, currency, freshness, etc.) or on a given knowledge-state (e.g., data source reputation, verifiability).

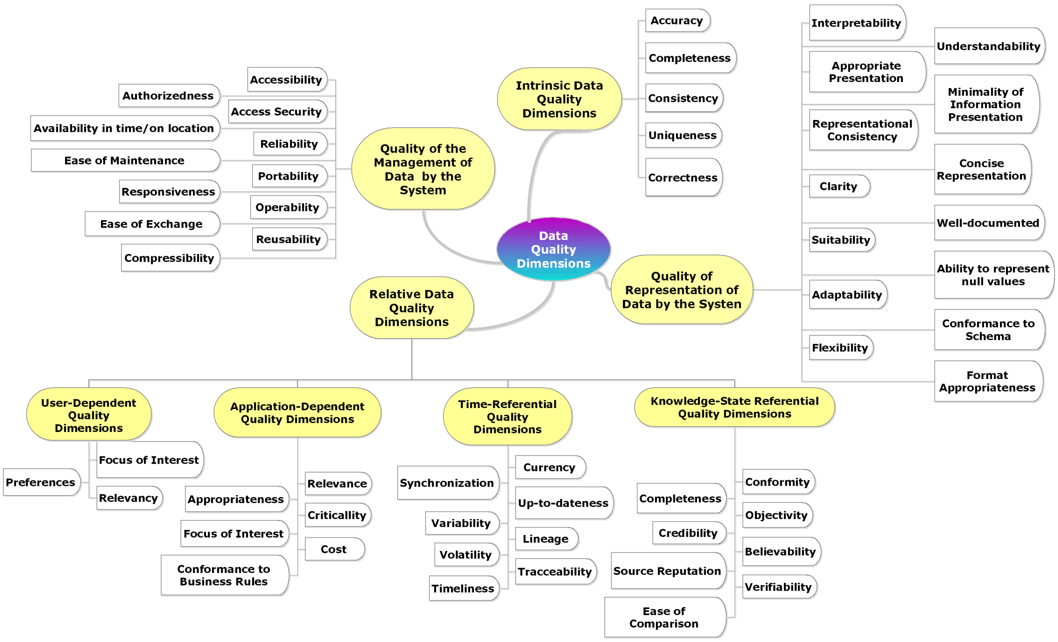


FIGURE 1. *Map of Data Quality Dimensions*

Data Quality Measurement Techniques

The statistical aspects of data quality have been the primary focus of statistical methods of imputation (*i.e.*, inferring missing data from statistical patterns of available data), predicting accuracy of the estimates based on the given data, data edits, automating detection and handling of outliers in data [9,17,40]. Utilization of statistical techniques for improving correctness of databases through introduction of new integrity constraints were proposed in [34]. The constraints are derived from the database instances using the conventional statistical techniques (e.g., sampling and regression), and every update of the database is validated against these constraints. If an update does not comply with them, then the data administrator is alerted and prompted to check correctness of the update.

Since databases model a portion of the real world which constantly evolves, the data quality estimates become outdated as time passes. Therefore, the estimation process should be repeated periodically depending on the dynamics of what is being modeled.

The general trend is the use of Artificial Intelligence methods (machine learning, knowledge representation schemes, management of uncertainty) for data validation [17,18,90]. The use of machine learning techniques for data validation and correction was first presented by Parsaye and Chignell: rules inferred from the database instances by machine learning methods were used to identify outliers in data and facilitate data validation process. Another similar approach was proposed by [73].

Exploratory Data Mining *(EDM)* [17] is a set of statistical techniques providing summaries that characterize data with typical values (medians and averages), variance, range, quantiles and correlations. Used as a first pass, EDM methods can be advantageously employed for data pre-processing before carrying out more expensive analyses. EDM aims to be widely applicable while dealing with unfamiliar datasets. These techniques have a quick response time, and have results which are easy to interpret, to store, and to update. Exploratory Data Mining can either be driven by the models to facilitate the use of parametric methods (model log-linear, for instance) or be driven by the data without any prior assumptions about inter-relationship between data using well-known non-parametric techniques for exploring multivariate distributions such as clustering, hierarchical or neural networks. The EDM summaries (e.g., averages, standard deviations, medians or other quantiles) can be used to characterize the data distribution, the correlations between attributes, or the center of the value distribution of a representative attribute, and also can be used to quantify and describe the dispersion of the values of the attribute around the center (form, density, symmetry, etc.). Other techniques are used to detect and cope with other problems on data such as missing values, improbable outliers and incomplete values. Concerning the techniques of analysis on missing data, the method of imputation through regression described by Little and Rubin [44] is generally used. Other methods such as Markov Chain Monte Carlo *(MCMC)* [72] are used to simulate data under the multivariate normal distribution assumption. Other references related to the problem of missing values are described in the tutorial by Pearson [60]. Concerning the isolated data (outliers): the techniques of detection employed are mainly control charts and various techniques based on *i)* a model, *ii)* on geometrical methods for distance measurement in the dataset (called geometric outliers), or *iii)* on the distribution (or the density) of data population [40] with the concept of local exception (called local distributional outliers) [9]. Other tests of "goodness-of-fit" such as *Chi2* check the independence of the attributes. The Kolmogorov-Smirnov test provides a measure of the maximum distance between the supposed distribution of the data and the empirical distribution computed from the data. These univariate tests are very useful to validate the analysis techniques and the assumptions on the used models. Moreover, complex and multivariate tests can be used such as pyramids, hyper-pyramids and Mahalanobis test for distances between multivariate averages [38]. The interested reader is invited to read the survey of Pyle [64], in particular for the use of entropy as a preliminary data characterization measure ([64], Section 11.3), and [17] for an illustrative description of these techniques.

Data Quality Models

In practice, assessing data quality in database systems has mainly been conducted by professional assessors with more and more cost-competitive auditing practices. Well-known approaches from industrial quality management and software quality assessment have been adapted for data quality and have proposed an extension of metadata management [68,87] for data quality. The use of metadata for data quality evaluation and improvement has been advocated by many authors [17,49,69]. Rothenberg argued that information producers should perform *Verification, Validation, and Certification* (*VV&C*) of their data and that they should provide data quality metadata along with the datasets [69]. Several propositions fully integrate the modeling and the management of quality metadata into database design. Among these process-oriented approaches, the *TDQM* program (*Total Data Quality Management*) proposed by Wang *et al.* at the *Massachusetts Institute of Technology* [84,86] provides a methodology including the modeling of data quality in the Entity-Relationship conceptual database model. It also proposes guidelines for adding step-by-step data quality metadata on each element of the model (entity, attribute, association).

Other works have taken other (still similar) approaches in modeling and capturing the quality of relational data [50,70] and of semi-structured data (e.g., D^2Q [71]). Figure 2 presents a generic model of data quality with UML formalism that synthesizes the approaches.

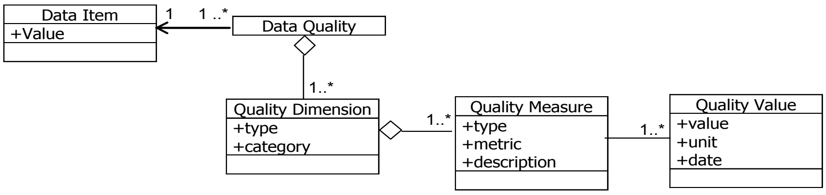


FIGURE 2. *Data Quality Model*

Most of the proposed data quality models rely on data quality metadata being available. Unfortunately, these approaches rely on precise and accurate metadata. However, such metadata are not always available and no commonly accepted standard describing data quality dimensions currently exists. Although considerable efforts have been invested in the development of metadata standard vocabularies for the exchange of information across different application domains (mainly for geographic information systems [30] and digital libraries) including substantial work on data quality characterization in theses domains, the obvious fact is that in practice the quality metadata in many application domains remains a luxury.

More specifically for data warehouse systems, many propositions concern the definition of quality models [36,37,80,83] with particular attention paid to data lineage and data transformation logs [16]. These metadata are very useful for the analysis and the interpretation of the probability distributions on the truncated or censured data, and also for debugging, implementing quality feedback loops and analyzing the causes of data errors.

Data Quality Methodologies

There has also been considerable work on methodologies for measuring, improving or assessing the quality of data in databases: e.g., *TDQM - Total Data Quality Management* – [25,86], Redman's methodology [68], IP-MAP [71]. These three methodologies were proposed for traditional management information systems (MIS), and not especially for data warehouses or heterogeneous cooperative systems [25]. All methodologies deal with both process-centric and data-centric activities. Process-centric activities propose to improve data quality, modifying some aspect of the processes that manage data and intra-organizational data flows. Data-centric activities improve data quality independently from the processes. Only the TDQM methodology provides guidelines for economic aspects related to data quality, specifically for evaluation of the cost of quality loss, the costs of the data improvement process and the benefits and savings resulting from data quality improvement (see Figure 3 for a comparative presentation). Redman's methodology [68] provides a huge number of case studies that provide evidence of the high cost related with low quality in information systems. A methodology specifically focused on costs and savings is described in [46]. Avenali *et al.* in [2] describe an algorithm for optimal choice of sources in terms of improvement of quality/cost balancing for a given demand of data and corresponding quality in a cooperative information system. However, such methodologies rely on human assessment of data, which is often time-consuming and possibly error-prone. Previous works have assumed that the metadata regarding the quality of data is available, accurate, and unbiased; either published by the data providers themselves or provided by user rankings of the most accurate or reliable data sources.

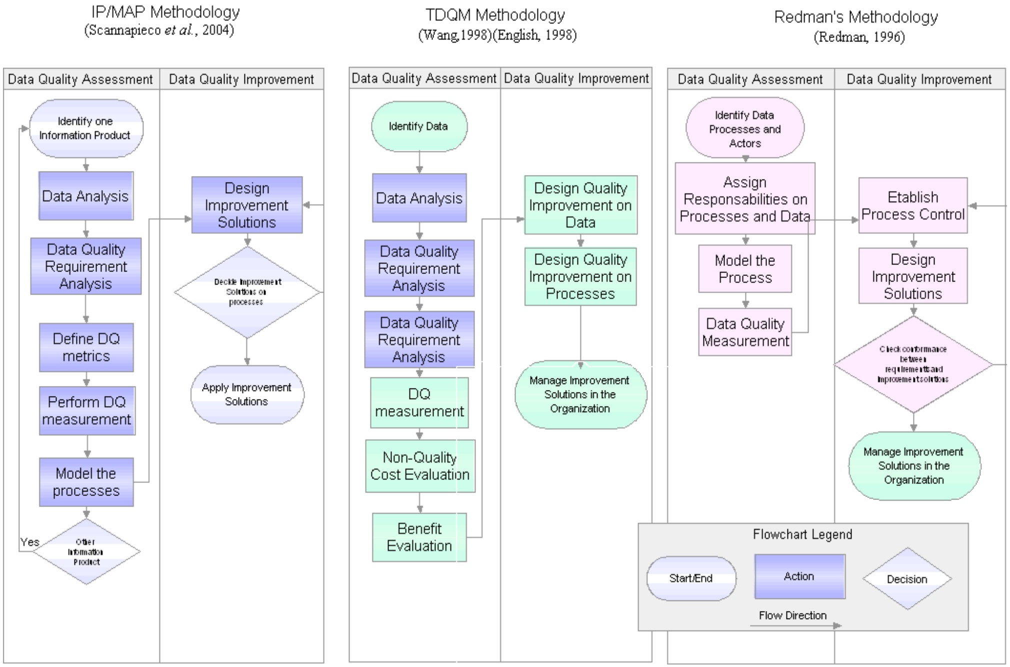

IP/MAP Methodology
(Scannapieco et al., 2004)
TDQM Methodology
(Wang,1998)(English, 1998)
Redman's Methodology
(Redman, 1996)
Data Quality Assessment
Data Quality Improvement
Identify one Information Product
Data Analysis
Data Quality Requirement Analysis
Define DQ metrics
Perform DQ measurement
Model the processes
Yes
Other Information Product
Design Improvement Solutions
Decide Improvement Solutions on processes
Apply Improvement Solutions
Identify Data
Data Analysis
Data Quality Requirement Analysis
Data Quality Requirement Analysis
DQ measurement
Non-Quality Cost Evaluation
Benefit Evaluation
Design Quality Improvement on Data
Design Quality Improvement on Processes
Manage Improvement Solutions in the Organization
Identify Data Processes and Actors
Assign Responsabilities on Processes and Data
Model the Process
Data Quality Measurement
Etablish Process Control
Design Improvement Solutions
Check conformance between requirements and improvement solutions
Manage Improvement Solutions in the Organization
Flowchart Legend
Start/End
Action
Decision
Flow Direction

FIGURE 3. *IP/MAP, TDQM and Redman's Methodology for Data Quality Management*

This overview is focused on the state of the art in data quality management with a wide description range of related works from data quality characterization and modeling (with quality dimension definitions, metrics and models) to higher level methodological approaches. To complete this overview, the next subsection will present the main techniques of data quality improvement, in particular record linkage and data cleaning techniques with Extraction-Transformation-Loading *(ETL)* operations.

3.2. Techniques for Data Cleaning

Record Linkage

The problem of detecting duplicate entities that describe the same real-world object (and purging them) is an important data cleaning task necessary to improve data quality, in particular in the context of data integration prior to warehousing, where data from distributed and heterogeneous data sources is combined and loaded into a data warehouse system. Deduplication of relational data received considerable attention in Database, Knowledge Discovery and Data Mining communities under various names: record linkage [27], merge/purge problem [33], object matching [15], etc. as mentioned in the introduction. Domain-dependent solutions were proposed for detecting duplicates for Census datasets [89], medical data [56], genealogical data [65] or bibliographic data [7,48]. Domain-independent solutions include identity uncertainty [59], entity identification [43], entity resolution [6], or approximate string joins [31]. More generally, most of the methods proposed in the literature for Record Linkage consist of the five following steps:

i) **Pre-processing** for coding, formatting and standardizing the data to compare;

ii) **Selecting a blocking method** to reduce the search space by partitioning the data file into mutually exclusive blocks to compare (e.g., hashing, sorting keys, sorted nearest neighbors or windowing over one or more keys (*i.e.*, attributes) [33]). The constraint on computational complexity is worth considering for the choice of the method: blocking function is $O(h(n)+n^2/b)$ with b the number of blocks with n records per block and $b.O(n^2/b^2)=O(n^2/b)$ the number of pairs of records to compare *;* hashing function (such as $h(n)=n$) is preferable rather than sorting function (such as $h(n)=n \log n$); Sorted Nearest-Neighbors is $O(n \log n+wn)$ for a window of size w. This method reduces the maximum number of comparisons per record to $2w-1$ and $(w-1)(n-w/2)$ is the number of pairs to compare by sorted NN*;*

iii) **Selecting and computing a comparison function**: this step consists in measuring the similarity distance between the pairs of records [55], e.g., using a simple string distance, or the occurrence frequency-weighted distance (see Table 2 for other distances such as Hamming distance, N-grams, Soundex, etc.);

iv) **Selecting a decision model**: this step consists in assigning and classifying pairs of records as matching, non-matching or potentially matching records with methods that can be probabilistic (with or without training datasets), knowledge-based or empirical (see Table 3);

v) **Validation of the method and feedback.**

Distances	Main Characteristics
Hamming Distance	Fixed numerical fields (e.g., SSN, Zip Code) without considering added/missing characters
Edit Distance	Compute the minimal cost of transformation (using add/drop/exchange of characters (Smith and Waterman)
Jaro's Algorithm	(C/L1 +C/L2 +(2C-T)/2C)/3 with C the number of common characters and T the number of transposed characters in two strings of length L1 and L2
N-grams Distance	$\sqrt{\sum_{\forall x} \lvert f_a(x) - f_b(x) \rvert}$ extended by Q-grams (Gravano *et al.*, VLBD 2001)
Soundex	Based on phonetics and consonants (e.g., the code of "John" and "Jon" is J500) with the Soundex Code Guide

Letters	*Number*
B, F, P, V	1
C, G, J, K, Q, S, X, Z	2
D, T	3
L	4
M, N	5
R	6

Table 2 – Main Characteristics of Distances Functions for String Matching

In the data cleaning literature, there exist various decisional models mainly used for record linking or duplicate identification (Table 3). The problem of identifying duplicate records in databases was originally identified by Newcombe *et al.* in 1959 [56] as record linkage on medical records for identifying the same individual over different time periods. Fellegi and Sunter [27] developed a formal theory for record linkage and offered statistical methods for estimating matching parameters and error rates. In more recent work in statistics, Winkler [88] proposed using Dempster *et al.*'s EM-based method [21] for obtaining optimal matching rules. Among the empirical approaches, Hernandez and Stolfo [33] developed the sorted neighborhood method for limiting the number of potential duplicate pairs that require distance computation.

Model *(Tool)*	Authors	Model
Error-based Model	Fellegi and Sunter, 1969 [27]	Probabilistic
EM-based Method	Dempster *et al.*, 1977, [21]	
Induction	Bilenko and Mooney, 2003 [6]	
Clustering for Record Linkage *(Tailor)*	Elfeky *et al.*, 2002, [24]	
1-1 matching and Bridging File	Winkler, 2003, [89]	
Sorted-Nearest Neighbors method	Hernandez and Stolfo, 1995, [33]	
XML Object Matching	Weis and Naumann, 2004, [88]	Empirical
Hierarchical Structure *(Delphi)*	Ananthakrishna *et al.*, 2002, [1]	
Matching Prediction based on clues	Buechi *et al.*, 2003, [11]	
Functional Dependencies Inference	Lim *et al.*,1993, [43]	Knowledge-Based
Transformation function *(Active Atlas)*	Tejada *et al.*,2001, [78]	
Rules and sorted-NN *(Intelliclean)*	(Low *et al.*, 2001, [47]	

Table 3 - Decision Models for Duplicate Identification and Record Linkag

Tejada *et al.* [78] developed a system that employs active learning methods for selecting record pairs that are informative for training the record-level classifier that combines similarity estimates from multiple fields across different metrics. In all of these approaches fixed-cost similarity metrics were used to compare the database records. Other approaches are reported in Table 3 and presented by [5].

Extraction, Transformation and Loading for Data Cleaning

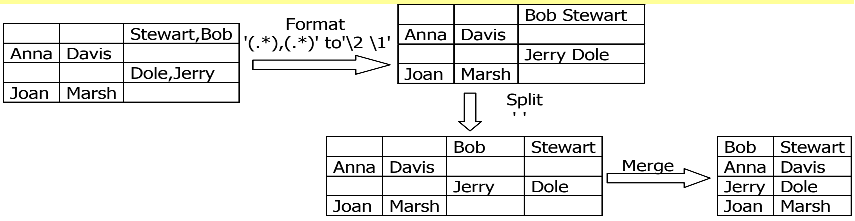
Under the general acronym *ETL*, the *Extraction-Transformation-Loading* activities cover the most prominent tasks of data preparation before the warehousing and mining processes [82]. They include: *i)* the identification of relevant information at the source side, *ii)* the extraction of this information, *iii)* the transformation and integration of the information coming from multiple sources into a common format and, *iv)* the cleaning and correction of the integrated dataset. Despite the specialized ETL tools (mainly dedicated to relational data) available in the market, data preparation and cleaning processes remain complex, costly and critical [75]. This area has raised lot of interest from the research community [66,82,89], now focusing on semi-structured data [88].

Several academic tools and algorithms were proposed for data transformation and conciliation: AJAX [29], Potter's Wheel [67], ARKTOS [81], Telcordia [13], and for Record Linkage: IntelliClean [47], Tailor [24], ClueMaker [11]. Table 4 presents the main ETL operators and Figure 4 gives a simple but illustrative example of data transformation with Potter’s Wheel operators [67].

TRANSFORMATION	DEFINITION
Potter's Wheel Format Ajax Map	*Applies a function to every value in an attribute column of a relational table (such as regular-expression based substitutions and arithmetic operations or user-defined functions).*
Add, Drop, Copy	*Allow users to add a new column, or to drop or copy a column.*
Merge	*Transforms concatenates values in two columns, optionally intrposing a constant in the middle, to form a new column*
Split	*Splits a column into two or more parts, and is used typically to parse a value into its constituent parts. The split positions can be specified by character positions, regular expressions, or by interactively performing splits on example values.*
Divide	*Conditionnaly divides a column, sending values into one of two new columns based on a predicate*
Fold	*Flattens tables by converting one row into multiple rows, folding a set of columns together into one column and replicating the rest. Conversely Unfold unflattens tables : it takes two columns , collects row that have the same values for all the other columns, and unfolds the two chosen columns.*

Table 4 – Main Data Transformation Operators for ETL

Example of data transformations: (Raman and Hellerstein, 2001)



		Stewart,Bob
Anna	Davis	
		Dole,Jerry
Joan	Marsh	

		Bob Stewart
Anna	Davis	
		Jerry Dole
Joan	Marsh	

		Bob	Stewart
Anna	Davis		
		Jerry	Dole
Joan	Marsh		

Bob	Stewart
Anna	Davis
Jerry	Dole
Joan	Marsh

Figure 4 – Example of Data Transformation [67]

AJAX [29] is an SQL extension for specifying each data transformation (such as matching, merging, mapping, and clustering) for the data cleaning process. These transformations standardize data formats when possible and find pairs of records that most probably refer to the same real-world object. The duplicate elimination step is applied if approximate duplicate records are found, and multi-table matching computes similarity joins between distinct data flows and consolidates them. Other propositions concern the definition of declarative language extensions for specifying/querying quality metadata or for applying data transformations necessary to specific cleaning process. The prototype *Q-Data* described by Sheth, Wood and Kashyap [74] checks if the existing data are correct and ensures data validation and cleanup by using a logical database language (*LDL++*). The system employs data validation constraints and data cleanup rules.

Both ETL tools and algorithms operate in a batch and off-line manner but "active data warehousing" (also called "real time warehousing") refers to a new trend where higher levels of data freshness are required for data warehouses that must be updated as frequently as possible with all the performance and overloading issues this raises for ETL tasks based on filters, transformers and binary operations over the multi-source datasets.

4. A Framework of Data Quality Awareness for KDD Process

Based on this exhaustive and database-oriented overview of data quality management, this section describes a pragmatic framework for data quality awareness preceding and during the knowledge discovery process. Each step may use and combine the previously mentioned approaches, methods and techniques proposed in the literature. The grand view of the framework is depicted in Figure 5. This framework is divided into parts upstream and downstream of the KDD process (respectively left-hand and right-hand sides of Figure 5). It consists of five steps from U1 to U5 for upstream and seven steps from D1 to D7 for downstream that will be described in the next subsections.

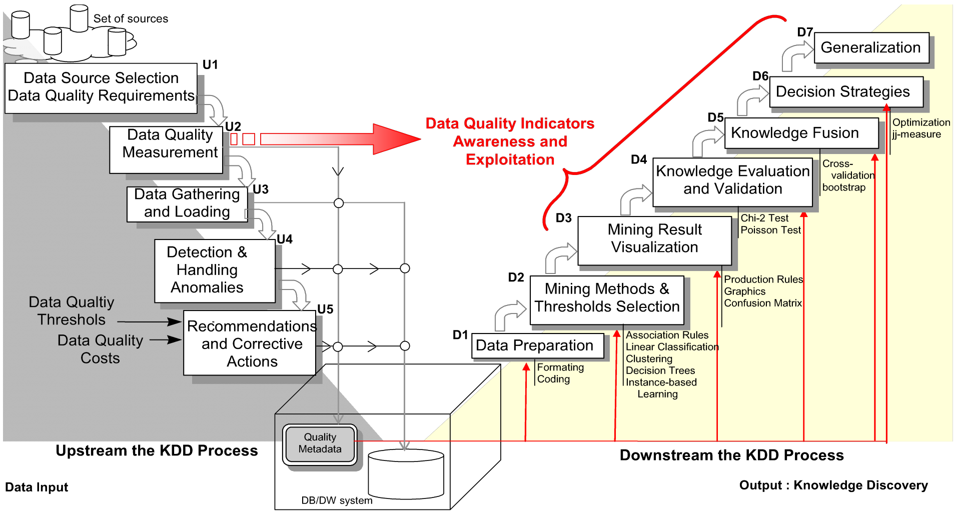
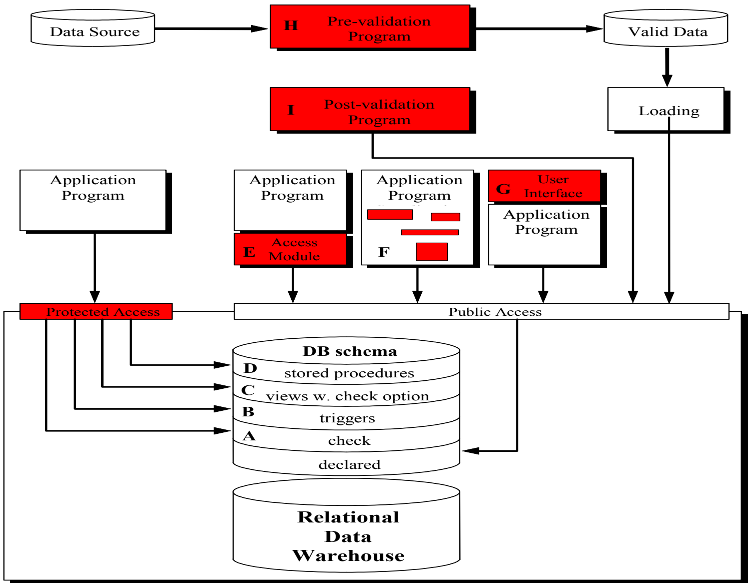


Figure 5 – General Framework for Data Quality Awareness in the KDD Process

4.1. KDD Upstream Quality-Aware Activities

The first upstream step denoted U1 in Figure 5 consists of *i)* selecting the data sources from which data will be extracted by automatic gathering and massive import procedures and *ii)* defining a clear, consistent set of data quality requirements for the data warehouse, for the input data sources and the entire KDD workflow and information chain (*i.e.*, qualitative and quantitative descriptions of quality criteria for information units, KDD processes and sub-processes).

In the U2 step, it is necessary to provide the basis for establishing measures and statistical controls of the data quality dimensions previously defined in the first step. Data items do not have the same importance, they may not be equivalent from a "strategic" point of view for the company and thus do not have to be considered in a uniform way for scheduling ETL and mining activities. The data quality measurement step U2 should provide a first characterization the data quality of the pre-selected data sources (e.g., by Exploratory Data Mining techniques) prior to data loading into the data warehouse. The EDM summaries will be stored as quality metadata characterizing the quality of each source, origin of the data. Different levels of periodic measurement and control can be implemented as shown in Figure 6 (listed



A.	Constraint predicates (e.g., SQL check, assertions)	**F.**	Distributed source code in application program
B.	Triggers SQL	**G.**	Validation procedure for data input forms
C.	Filtering views (SQL views with check option)	**H.**	Validation program before data loading
D.	Stored procedures associated to the SQL schema	**I.**	Validation program after data loading
E.	Access modules	**J.**	Manual validation

from **A.** to **J.**). The $\boxed{\text{U2}}$ step consists in computing quality scores and indicators mainly using pre-validation programs (see **H.** in Figure 6).
The $\boxed{\text{U3}}$ step consists of possibly cleaning, reconciling, aggregating data and loading data in to data warehouse with appropriate ETL tools and record linking strategies. The goal of the $\boxed{\text{U4}}$ step is to detect the problems of data quality (errors, outliers or poor quality data) throughout data processing with post-validation programs on the data warehouse system (see **I.** in Figure 6), to refresh the quality metadata and also to analyze the causes depending on the data quality (non-)tolerance thresholds and data quality costs. The purpose of the $\boxed{\text{U5}}$ step is to propose corrective actions and recommendations of improvements for each of the four previous upstream steps in order to set up quality feedback loops.

FIGURE 6 – *Different Levels for Controlling and Measuring Data Quality*

4.2. KDD Downstream: Quality Metadata Exploitation

Starting the KDD process, the classical step of data preparation $\boxed{\text{D1}}$ in Figure 5 consists of a succession of tasks such as *i)* selection of datasets and objects, *ii)* selection and weighting of the variables and features, *iii)* (re-)coding data, *iv)* analyzing missing values and their sources: different kinds of missing values are distinguished (e.g., unknown vs. unrecorded vs. irrelevant vs. results from censuring or truncating), *v)* detection of data anomalies: because outliers are individual data values that are inconsistent with the majority of values in the data collection, it's important to characterize the side-effects of the strategy of detecting and omitting outliers, *vi)* the homogenization of the data files, *vii)* the discretization of the

(1)

continuous attributes and *viii)* the possible use of quantization strategies for real variables (e.g., defining quartiles or deciles) or high-dimension reduction techniques.

The next steps $\boxed{\text{D2}}$ and $\boxed{\text{D3}}$ in Figure 5 consist of selecting the mining methods, configuration parameters and thresholds and, therefore the knowledge representation for visualizing the mining results (e.g., decision tables, decision trees, classification rules, association rules, instance-based learning representation and clusters). For quality-awareness, these steps (in particular $\boxed{\text{D3}}$) should include the visualization of data quality indicators previously computed in step $\boxed{\text{U2}}$ and stored in the quality metadata repository of the data warehouse.

The added-value of quality metadata consists in their exploitation as explanation factors for evaluating discovered knowledge and for validating one mining process (step $\boxed{\text{D4}}$) and combining the results of several mining techniques (step $\boxed{\text{D5}}$) and also drive the decision strategies (step $\boxed{\text{D6}}$) and generalization (step $\boxed{\text{D7}}$).

The next section will focus mainly on steps $\boxed{\text{U2}}$, $\boxed{\text{D3}}$ and $\boxed{\text{D4}}$ of Figure 5 in the context of association rule discovery and will present formally how quality metadata awareness and exploitation can be used to predict the cost of low-quality data over association rule discovery.

4.3. Cost of Low-Quality Data on Association Rules Discovery

Our initial assumption is that the quality of an association rule depends on the quality of the data which the rule is computed from. This section will present the formal definitions for introducing data quality indicators and how they can be combined as quality indicators for association rules.

Formal Definitions

Let *I* be a superset of items. An association rule *R* is an implication of the form:

LHS → *RHS* where $LHS \square I$, $RHS \square\ I$ and $LHS \square RHS = \square$. *LHS* and *RHS* are conjunctions of variables such as the extension of *LHS* is $g(LHS)= x_1 \square x_2 \square \ldots \square x_n$ and the extension of *Y* is $g(RHS)= y_1 \square y_2 \square \ldots \square y_{n'}$.

Let *j* (*j=1, 2,…, k*) be the dimensions of data quality (e.g., data completeness, freshness, accuracy, credibility, etc.). Let $q_{ij} \square [min_{ij} , max_{ij}]$ be a scoring value for the dataset I_i on the quality dimension *j* ($I_i \square I$). The vector, that keeps the values of all quality dimensions for each data item (normalized in [0,1]) is the called quality vector noted $q(I_i)$. The set of all possible quality vectors is called quality space *Q*.

Definition 1. Association Rule Quality

The quality of the association rule *R* is defined by the fusion function denoted "○" that merges the quality vectors of each item set constituting the extension of the right-hand and left-hand sides of the rule *R* such as:

$$Quality(R) = q(LHS) \circ q(RHS)$$

$$= q(x_1) \circ q(x_2) \circ \ldots \circ q(x_n) \circ q(y_1) \circ q(y_2) \circ \ldots \circ q(y_{n'})$$

Definition 2. Quality Score Fusion

Let T be the domain of values of the quality score $q_j(I_i)$ (also denoted q_{ij}) for the dataset I_i on the quality dimension j. The fusion function denoted "○" is commutative and associative such as ○: $T \cdot T \rightarrow T$. The fusion function may have different definitions depending on the considered quality dimension in order to suit the properties of each quality criterion. Table 5 presents several possible definitions for the fusion functions allowing the aggregation of quality scores per quality dimension for two datasets noted *x* and y over four considered quality dimensions.

j	**QUALITY DIMENSION**	**FUSION FUNCTION "○"**	**QUALITY DIMENSION S OF THE RULE $x \rightarrow y$**
1	Freshness	$\min[q_1(x),q_1(y)]$	The freshness of the association rule $x \rightarrow y$ is estimated pessimistically as the lower score of freshness.
2	Accuracy	$q_2(x).\ q_2(y)$	The accuracy of the association rule $x \rightarrow y$ is estimated as the probability of accuracy of the two data sets x and y.
3	Completeness	$q_3(x)+ q_3(y)- q_3(x)\ .\ q_3(y)$	The completeness of the association rule $x \rightarrow y$ is estimated as the probability that one of the two data sets is complete.
4	Consistency	$\max[q_4(x),\ q_4(y)]$	The consistency of the association rule $x \rightarrow y$ is estimated optimistically as the higher score of consistency.

Table 5. Fusion Functions for Merging Quality Scores per Dimension

These data quality indicators may be computed in step $\boxed{\text{U2}}$ of the framework (Figure 5) and may be flexibility aggregated and combined to give quality information on the discovered association rules (step $\boxed{\text{D3}}$) for mining results visualization.

Cost-Based Probabilistic Model

Despite good confidence, support or other interestingness measures, selecting an association rule is a decision that designates the rule as legitimately interesting (noted D_1), potentially interesting (D_2), or not interesting (D_3) based on the information contained in the quality vectors of the data item sets composing the *LHS* and *RHS* parts of the rule.

Consider the item *x* □ *LHS* □ *RHS* of a given association rule, let us use $P_{CE}(x)$ to denote the probability that the item *x* will be classified as “erroneous” (or “polluted”) with reference to one or more quality dimensions relevant to the application (e.g., freshness, accuracy, etc.), and $P_{CC}(x)$ denotes the probability that the item *x* will be classified as “correct” (*i.e.*, in the range of acceptable values for each pre-selected quality dimension). Also, $P_{AE}(x)$ represents the probability that the item *x* is actually erroneous (*AE*), and $P_{AC}(x)$ represents the probability that it is actually correct (*AC*) (see Figure 7).

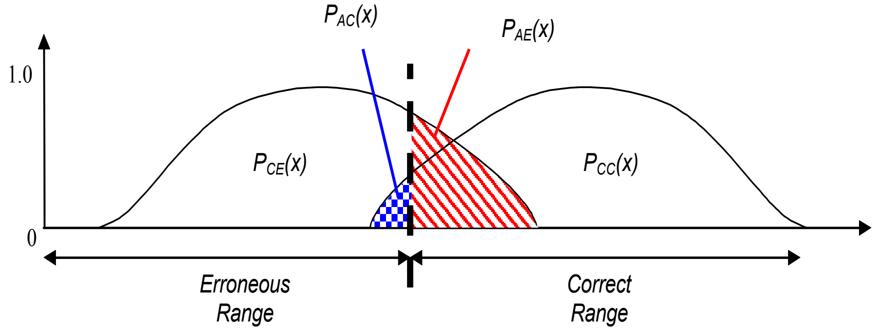


FIGURE 7. *Probabilities of detection of correct or polluted data*

The quality dimensions of x are measured and aggregated. For an arbitrary average quality vector $\overline{q}$□Q on all data items in *LHS* □*RHS* of the rule, we denote by $P(\overline{q}$□$Q \mid CC)$ or $f_{CC}(\overline{q})$ the conditional probability of the pattern $\overline{q}$ that corresponds to the average of quality vectors of the items that are classified as correct (*CC*). Similarly, we denote by $P(\overline{q}$□$Q \mid CE)$ or $f_{CE}(\overline{q})$ the conditional probability of the pattern $\overline{q}$ corresponds to the average of quality vectors of the items that are classified erroneous (*CE*). We denote by d the decision of the rule classification (*i.e.*, legitimately interesting D_1, potentially interesting D_2, or not interesting D_3), and by s the actual status of quality of the item sets upon which the rule has been computed. Let us also denote by $P(d=D_i, s=j)$ and $P(d=D_i \mid s=j)$ correspondingly, the joint and the conditional probability that the decision D_i is taken when the actual status of data quality (*i.e.*, *CC, CE, AE, AC*) is j. We also denote by c_{ij} the cost of making a decision D_i for classifying an association rule with actual data quality status j of the datasets composing the parts of the rule. Based on the example in Table 6 where we can see how the cost of decisions could affect the result of selection among interesting association rules, we need to minimize the mean cost c that results from making such a decision.

Cost	***Decision for Rule Selection***	***Actual Data Quality Status***
c_{10}	D_1	*CC*
c_{11}	D_1	*CE*
c_{12}	D_1	*AE*
c_{13}	D_1	*AC*
c_{20}	D_2	*CC*
c_{21}	D_2	*CE*
c_{22}	D_2	*AE*
c_{23}	D_2	*AC*
c_{30}	D_3	*CC*
c_{31}	D_3	*CE*
c_{32}	D_3	*AE*
c_{33}	D_3	*AC*

Table 6. Costs of various decisions for classifying association rules

(4)(3)

(2)
(6)

(5)

The corresponding mean cost is written as follows:

$$\begin{aligned}\bar{c} = {} & c_{10}.P(d = D_1, s = CC) + c_{20}.P(d = D_2, s = CC) + c_{30}.P(d = D_3, s = CC) \\ & + c_{11}.P(d = D_1, s = CE) + c_{21}.P(d = D_2, s = CE) + c_{31}.P(d = D_3, s = CE) \\ & + c_{12}.P(d = D_1, s = AE) + c_{22}.P(d = D_2, s = AE) + c_{32}.P(d = D_3, s = AE) \\ & + c_{13}.P(d = D_1, s = AE) + c_{23}.P(d = D_2, s = AE) + c_{33}.P(d = D_3, s = AC)\end{aligned}$$

From the Bayes theorem, the following is true:

$$P(d = D_i, s = j) = P(d = D_i \mid s = j).P(s = j)$$

where *i=1,2,3* and *j= CC,CE,AE,AC*. Let us also assume that $\overline{q}$ is the average quality vector drawn randomly from the space of all quality vectors of the item sets of the rule. The following equality holds for the conditional probability $P(d=D_i| s=j)$:

$$P(d = D_i \left| s = j \right.) = \sum_{q \in Q_i} f_j(\overline{q}).$$

where *i=1,2,3* and *j=CC,CE,AE,AC*. f_j is the probability density of the quality vectors when the actual quality status is *j*. We also denote the a priori probability of *CC* or *P(s=CC)* as π^0, the a priori probability of $P(s=AC)=\pi^0{}_{AC}$, the a priori probability of $P(s=AE)=\pi^0{}_{AE}$ and the a priori probability of $P(s=CE)=1- \pi^0+ \pi^0{}_{AE} - \pi^0{}_{AC}$. Without the misclassification region *P(s=CE)* could be simplified as $1- \pi^0$. The mean $\bar{c}$ost c in Eq. (2) based on Eq. (3) is written as follows:

$$\begin{aligned}\bar{c} = {} & c_{10}.P(d = D_1 \left| s = CC \right.).P(s = CC) + c_{20}.P(d = D_2 \left| s = CC \right.).P(s = CC) \\ & + c_{30}.P(d = D_3 \left| s = CC \right.).P(s = CC) + c_{11}.P(d = D_1 \left| s = CE \right.).P(s = CE) \\ & + c_{21}.P(d = D_2 \left| s = CE \right.).P(s = CE) + c_{31}.P(d = D_3 \left| s = CE \right.).P(s = CE) \\ & + c_{12}.P(d = D_1 \left| s = AE \right.).P(s = AE) + c_{22}.P(d = D_2 \left| s = AE \right.).P(s = AE) \\ & + c_{32}.P(d = D_3 \left| s = AE \right.).P(s = AE) + c_{13}.P(d = D_1 \left| s = AC \right.).P(s = AC) \\ & + c_{23}.P(d = D_2 \left| s = AC \right.).P(s = AC) + c_{33}.P(d = D_3 \left| s = AC \right.).P(s = AC)\end{aligned}$$

and by using Eq. (4) and dropping the dependent vector variable $\overline{q}$, Eq. (5) becomes:

$$\begin{aligned}\bar{c} = {} & \sum_{q \in Q_1} \left[f_{CC}.c_{10}.\pi^0 + f_{CE}.c_{11}.(1 - \pi^0 - \pi^0_{AC} + \pi^0_{AE}) + f_{AE}.c_{12}.\pi^0_{AE} + f_{AC}.c_{13}.\pi^0_{AC} \right] \\ & + \sum_{q \in Q_2} \left[f_{CC}.c_{20}.\pi^0 + f_{CE}.c_{21}.(1 - \pi^0 - \pi^0_{AC} + \pi^0_{AE}) + f_{AE}.c_{22}.\pi^0_{AE} + f_{AC}.c_{23}.\pi^0_{AC} \right] \\ & + \sum_{q \in Q_3} \left[f_{CC}.c_{30}.\pi^0 + f_{CE}.c_{31}.(1 - \pi^0 - \pi^0_{AC} + \pi^0_{AE}) + f_{AE}.c_{32}.\pi^0_{AE} + f_{AC}.c_{33}.\pi^0_{AC} \right]\end{aligned}$$

Every point $\overline{q}$ in the quality space *Q* belongs to the partitions Q_1 or Q_2 or Q_3 that correspond respectively to the decision space *D* partitions: D_1 or in D_2 or D_3 in such a way that its contribution to the mean cost is minimum. This will lead to the optimal selection for the three sets of rules which we denote by $D^0{}_1$, $D^0{}_2$ and $D^0{}_3$. Based on this observation, a point *q* is assigned to the three optimal areas as follows:

(8)(9)(10)

To D_1^0 if :

$$f_{CC}.c_{10}.\pi^0 + f_{CE}.c_{11}.(1-\pi^0-\pi_{AC}^0+\pi_{AE}^0) + f_{AE}.c_{12}.\pi_{AE}^0 + f_{AC}.c_{13}.\pi_{AC}^0$$
$$\leq f_{CC}.c_{30}.\pi^0 + f_{CE}.c_{31}.(1-\pi^0-\pi_{AC}^0+\pi_{AE}^0) + f_{AE}.c_{32}.\pi_{AE}^0 + f_{AC}.c_{33}.\pi_{AC}^0$$
$$\text{and } f_{CC}.c_{10}.\pi^0 + f_{CE}.c_{11}.(1-\pi^0-\pi_{AC}^0+\pi_{AE}^0) + f_{AE}.c_{12}.\pi_{AE}^0 + f_{AC}.c_{13}.\pi_{AC}^0$$
$$\leq f_{CC}.c_{20}.\pi^0 + f_{CE}.c_{21}.(1-\pi^0-\pi_{AC}^0+\pi_{AE}^0) + f_{AE}.c_{22}.\pi_{AE}^0 + f_{AC}.c_{23}.\pi_{AC}^0$$

To D_2^0 if :

$$f_{CC}.c_{20}.\pi^0 + f_{CE}.c_{21}.(1-\pi^0-\pi_{AC}^0+\pi_{AE}^0) + f_{AE}.c_{22}.\pi_{AE}^0 + f_{AC}.c_{23}.\pi_{AC}^0$$
$$\leq f_{CC}.c_{30}.\pi^0 + f_{CE}.c_{31}.(1-\pi^0-\pi_{AC}^0+\pi_{AE}^0) + f_{AE}.c_{32}.\pi_{AE}^0 + f_{AC}.c_{33}.\pi_{AC}^0$$
$$\text{and } f_{CC}.c_{20}.\pi^0 + f_{CE}.c_{21}.(1-\pi^0-\pi_{AC}^0+\pi_{AE}^0) + f_{AE}.c_{22}.\pi_{AE}^0 + f_{AC}.c_{23}.\pi_{AC}^0$$
$$\leq f_{CC}.c_{10}.\pi^0 + f_{CE}.c_{11}.(1-\pi^0-\pi_{AC}^0+\pi_{AE}^0) + f_{AE}.c_{12}.\pi_{AE}^0 + f_{AC}.c_{13}.\pi_{AC}^0$$

To D_3^0 if :

$$f_{CC}.c_{30}.\pi^0 + f_{CE}.c_{31}.(1-\pi^0-\pi_{AC}^0+\pi_{AE}^0) + f_{AE}.c_{32}.\pi_{AE}^0 + f_{AC}.c_{33}.\pi_{AC}^0$$
$$\leq f_{CC}.c_{10}.\pi^0 + f_{CE}.c_{11}.(1-\pi^0-\pi_{AC}^0+\pi_{AE}^0) + f_{AE}.c_{12}.\pi_{AE}^0 + f_{AC}.c_{13}.\pi_{AC}^0$$
$$\text{and } f_{CC}.c_{30}.\pi^0 + f_{CE}.c_{31}.(1-\pi^0-\pi_{AC}^0+\pi_{AE}^0) + f_{AE}.c_{32}.\pi_{AE}^0 + f_{AC}.c_{33}.\pi_{AC}^0$$
$$\leq f_{CC}.c_{20}.\pi^0 + f_{CE}.c_{21}.(1-\pi^0-\pi_{AC}^0+\pi_{AE}^0) + f_{AE}.c_{22}.\pi_{AE}^0 + f_{AC}.c_{23}.\pi_{AC}^0$$

For the sake of simplicity, let's now consider the case of the absence of the misclassification region (*i.e.*,f_{AC}, f_{AE} are null and $\pi^0{}_{AE} = \pi^0{}_{AC} = 0$, we thus can simplify the inequalities above:

$$D_1^0 = \left\{ \bar{q} : \frac{f_{CE}}{f_{CC}} \leq \frac{\pi^0}{1-\pi^0} \cdot \frac{c_{30}-c_{10}}{c_{11}-c_{31}} \text{ and, } \frac{f_{CE}}{f_{CC}} \leq \frac{\pi^0}{1-\pi^0} \cdot \frac{c_{20}-c_{10}}{c_{11}-c_{21}} \right\}$$
$$D_2^0 = \left\{ \bar{q} : \frac{f_{CE}}{f_{CC}} \geq \frac{\pi^0}{1-\pi^0} \cdot \frac{c_{20}-c_{10}}{c_{11}-c_{21}} \text{ and, } \frac{f_{CE}}{f_{CC}} \leq \frac{\pi^0}{1-\pi^0} \cdot \frac{c_{30}-c_{20}}{c_{21}-c_{31}} \right\}$$
$$D_3^0 = \left\{ \bar{q} : \frac{f_{CE}}{f_{CC}} \geq \frac{\pi^0}{1-\pi^0} \cdot \frac{c_{30}-c_{10}}{c_{11}-c_{31}} \text{ and, } \frac{f_{CE}}{f_{CC}} \geq \frac{\pi^0}{1-\pi^0} \cdot \frac{c_{30}-c_{20}}{c_{21}-c_{31}} \right\}$$

The inequalities of Eq. (10) give rise to three different threshold values *L, P* and *N* (respectively for legitimately, potentially and not interesting rules) in the decision space. Eq. (11) defines concretely the decision regions based on the cost of rule selection decision such as:

(11)

$$L = \frac{\pi^0}{1-\pi^0} \cdot \frac{c_{30} - c_{10}}{c_{11} - c_{31}}$$

$$P = \frac{\pi^0}{1-\pi^0} \cdot \frac{c_{20} - c_{10}}{c_{11} - c_{21}}$$

$$N = \frac{\pi^0}{1-\pi^0} \cdot \frac{c_{30} - c_{20}}{c_{21} - c_{31}}$$

These thresholds can be used then as predictive means for quality awareness in mining association rules for selecting legitimately interesting rules based on the data quality indicators.

5. Conclusion

This chapter gave an exhaustive overview of data quality management and related techniques that can be employed for improving the data quality awareness of knowledge discovery and data mining techniques. Three application examples have introduced the importance of data quality-awareness for knowledge discovery activities. The chapter also provided a pragmatic framework for quality-driven KDD process. Finally, a prospective work on a theoretical probabilistic cost model for estimating the cost of low-quality data in mining association rules based on data quality indicators has been presented. The future plans regarding this work, are to study the optimality of the decision model, to propose error estimation and to validate the model with experiments on large real datasets for selecting high-quality association rules with additional and confirmative combinations of data quality indicators.

References

1. Anathakrishna R., Chaudhuri S., Ganti V., Eliminating Fuzzy Duplicates in Datawarehouses, *Proc. of Intl. Conf. VLDB*, 2002.
2. Avenali A., Batini C., Bertolazzi P. , Missier P., A Formulation of the Data Quality Optimization Problem, *Proc. of the Intl. CAiSE Workhop on Data and Information Quality (DIQ)*, Riga, 2004.
3. Ballou D.P., Pazer H. Designing Information Systems to Optimize the Accuracy-Timeliness Trade-off, *Information Systems Research*, 6(1), 1995.
4. Ballou D.P., Pazer H., Modeling Completeness Versus Consistency Tradeoffs in Information Decision Contexts, *IEEE TKDE*, 15(1): 240-243, 2002.
5. Batini C., Catarci T. and Scannapiceco M., A survey of Data Quality Issues in Cooperative Information Systems, *Tutorial presented at Intl. Conf. on Conceptual Modeling (ER)*, 2004.
6. Bejelloun O., Garcia-Molina H., Su Q., Widom J., Swoosh: A generic Approach to Entity Resolution, Tech. Rep., Stanford Database Group, March 2005.
7. Bilenko M. and Raymond J. Mooney R.J., Adaptive Duplicate Detection Using Learnable String Similarity Measures, *Proc. of Intl. Conf. KDD*, p39-48, 2003.
8. Bouzeghoub M., Peralta V., A Framework for Analysis of Data Freshness, *Proc. of the ACM SIGMOD Workshop on Information Quality in Information Systems (IQIS)*, 2004.

9. Breunig M., Kriegel H., Ng R., Sander J., LOF: Identifying Density-Based Local Outliers, *Proc. of ACM SIGMOD Conf.*, p. 93-104, 2000.
10. Brodie M. L., Data Quality in Information Systems, *Information and Management*, vol. 3, p. 245-258, 1980.
11. Buechi M., Borthwick A., Winkel, A. and Goldberg A., ClueMaker: a Language for Approximate Record Matching, *Proc. of Inl. Conf. on Information Quality (ICIQ)*, Boton, MA, 2003.
12. Carlson D., Data Stewardship in Action, *DM Review*, May 2002.
13. Caruso F., Cochinwala M., Ganapathy U., Lalk G., Missier P., Telcordia's Database Reconciliation and Data Quality Analysis Tool, *Proc. of Intl. Conf. VLDB*, p. 615-618, 2000.
14. Celko J., McDonald J., Don't warehouse dirty data, *Datamation*, 41(18), 1995.
15. Chaudhuru S., Ganjam K., Ganti V., and Motwani R., Robust and efficient Fuzzy Match for Online Data Cleaning, *Proc. of ACM SIGMOD Conf.*, 2003.
16. Cui Y., Widom J., Lineage Tracing for General Data Warehouse Transformation, *Proc. of Intl. Conf. on VLDB*, p. 471-480, 2001.
17. Dasu T., Johnson T., Exploratory Data Mining and Data Cleaning, Wiley, 2003.
18. Dasu T., Johnson T., Muthukrishnan S., Shkapenyuk V., Mining Database Structure or, How to Build a Data Quality Browser, *Proc. of ACM SIGMOD Conf.*, 2002.
19. De Giacomo G., Lembo D., Lenzerini M., Rosati R., Tackling Inconsistencies in Data Integration through Source Preferences, *Proc. of the ACM SIGMOD Workshop on Information Quality in Information Systems (IQIS)*, 2004.
20. Delen G., Rijsenbrij D., The Specification, Engineering and Measurement of Information Systems Quality, *Journal of Software Systems*, no.17, p. 205-217, 1992.
21. Dempster A.P., Laird N.M. and Rubin D.B., Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society*, vol. 39, p. 1-38, 1977.
22. Djeraba C., Association and Content-Based Retrieval, *IEEE TDKE, vol. 15, n° 1*, p.118-135, 2003.
23. DuMouchel W., Volinsky C., Johnson T., Cortez C., Pregibon D., Squashing Flat Files Flatter, *Proc. of Intl. KDD Conf.*, p. 6-16, 1999.
24. Elfeky M.G., Verykios V.S., Elmagarmid A.K., Tailor: A Record Linkage Toolbox, *Proc. of the ICDE Conf.*, 2002.
25. English L., Improving Data Warehouse and Business Information Quality, Wiley, 1998.
26. Fan K., Lu H., Madnick S., Cheung D., Discovering and Reconciling Value Conflicts for Numerical Data Integration, *Information Systems*, vol.26, no.8, 2001.
27. Fellegi I.P., Sunter A.B., A Theory for Record Linkage, *Journal of the American Statistical Association*, 64, p. 1183-1210, 1969.
28. Fox C., Levitin A., Redman T., The Notion of Data and its Quality Dimensions, *Information Processing and Management*, vol. 30, no. 1, 1994.
29. Galhardas H., Florescu D., Shasha D., and Simon E., Saita C., Declarative Data Cleaning: Language, Model and Algorithms, *Proc. of Intl. Conf. VLDB*, p.371-380, 2001.
30. Goodchild M., Jeansoulin R. (Ed.), Data Quality in Geographic Information: From Error to Uncertainty, Hermès, 1998.
31. Gravano L., Ipeirotis P.G., Koudas N., and Srivastava D., Text Joins in an RDBMS for Web Data Integration, *Proc. of Intl. World Wide Web Conf. (WWW)*, 2003.
32. Guérin E., Marquet G., Burgun A., Loréal O., Berti-Equille L., Leser U., Moussouni F., Integrating and Warehousing Liver Gene Expression Data and Related Biomedical Resources in GEDAW, *Proc. of the 2nd Intl. Workshop on Data Integration in the Life Science (DILS)*, San Diego, 2005.
33. 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.

34. Hou W.C., Zhang Z., Enhancing Database Correctness: A Statistical Approach, *Proc. of Intl. Conf. ACM SIGMOD*, 1995.
35. Huang K., Lee Y., Wang R., Quality Information and Knowledge Management, Prentice Hall, New Jersey, 1999.
36. Jarke M., Jeusfeld M. A., Quix C., Vassiliadis P., Architecture and Quality in Data Warehouses, *Proc. of Intl. Conf. CAiSE*, p. 93-113, 1998.
37. Jeusfeld M.A., Quix C., Jarke M., Design and Analysis of Quality Information for Data Warehouses, *Proc. of Intl. Conf. Conceptual Modelling (ER)*, p. 349-362, 1998.
38. Johnson T., Dasu T., Comparing Massive High-Dimensional Data Sets, *Proc. of Intl. Conf. KDD*, p. 229-233, 1998.
39. Kahn B., Strong D., Wang R., Information Quality Benchmark: Product and Service Performance, *Com. of the ACM*, vol. 45, no.4, 2002.
40. Knorr E., Ng R., Algorithms for Mining Distance-Based Outliers in Large Datasets, *Proc. of Intl. Conf. VLDB*, p. 392-403, 1998.
41. Lavrač N., Flach P.A., and Zupan B., Rule Evaluation Measures: A Unifying View, *ILP*, p. 174-185, 1999.
42. Liepins G., Uppuluri V., Data Quality Control: Theory and Pragmatics, M. Dekker, 1990.
43. Lim L., Srivastava J., Prabhakar S., Richardson J., Entity Identification in Database Integration, *Proc. of Intl. ICDE Conf.*, 1993.
44. Little R.J., Rubin D.B., Statistical Analysis with Missing Data, Wiley, New-York, 1987.
45. Liu and Chi, Evolutionary Data Quality, *Proc. of Intl. Conf on Information Quality (ICIQ)*, 2002.
46. Loshin D., Enterprise Knowledge Management: The data quality approach, Morgan Kaufmann, 2001.
47. Low W.L., Lee M.L., Ling T.W., A Knowledge-Based Approach for Duplicate Elimination in Data Cleaning, *Information System*, Vol. 26 (8), 2001.
48. McCallum A., Nigam K., Ungar L.H., Efficient Clustering of High-Dimensional Data Sets with Application to Reference Matching, *Proc. of ACM KDD*, 2000.
49. Mihaila G. A., Raschid L., and Vidal M., Using Quality of Data Metadata for Source Selection and Ranking, *Proc. of the Intl. WebDB Workshop*, p. 93-98, 2000.
50. Missier P., Batini C., A Multidimensional Model for Information Quality in CIS, *Proc. of the Intl. Conf. on Information Quality (ICIQ)*, MIT, Boston, 2003
51. Monge A., Matching Algorithms Within a Duplicate Detection System, *IEEE Data Eng. Bull.*, 23(4):14-20, 2000.
52. Müller H., Leser U., Freytag J.C., Mining for Patterns in Contradictory Data, *Proc. of the ACM SIGMOD Workshop on Information Quality in Information Systems (IQIS)*, 2004.
53. Naumann F., Leser U., Freytag J., Quality-Driven Integration of Heterogeneous Information Systems, *Proc. of Intl. Conf. VLDB*, 1999.
54. Naumann F., Quality-Driven Query Answering for Integrated Information Systems, LNCS 2261, Springer 2002.
55. Navarro G., A Guided Tour to Approximate String Matching, *ACM Computer Surveys*, 33(1):31-88, 2001.
56. Newcombe, H. B., Kennedy, J. M. Axford, S. J., and James, A. P., Automatic Linkage of Vital Records, *Science*, 130, p. 954-959, 1959.
57. Ordonez C., Omiecinski E., Discovering Association Rules Based on Image Content, *IEEE Advances in Digital Libraries Conf. (ADL'99)*, 1999, p. 38-49.
58. Paradice D.B., Fuerst W.L., A MIS Data Quality Management Strategy Based on an Optimal Methodology. *Journal of Information Systems*, 5(1):48 - 66, 1991.
59. Pasula H., Marthi B., Milch B., Russell S., and Shpitser I., Identity Uncertainty and Citation Matching, In *Advances in Neural Information Processing Systems*, MIT Press, 2003.
60. Pearson R.K., Data Mining in Face of Contaminated and Incomplete Records, *Proc. of SIAM Intl. Conf. Data Mining*, 2002.

61. Perner P., Data Mining on Multimedia, LNCS vol. 2558, Springer, 2002.
62. Piattini M., Genero M., Calero C., Polo C., Ruiz F., Database Quality, In Chapter 14: Advanced Database Technology and Design, Artech House, p. 485-509, 2000.
63. Piattini, M., Calero C. Genero M. (Eds.), Information and Database Quality, The Kluwer International Series on Advances in Database Systems, Vol. 25, 2002.
64. Pyle D., Data Preparation for Data Mining, Morgan Kaufmann, 1999.
65. Quass D. And Starkey P., Record Linkage for Genealogical Databases, Proc. of Intl. KDD Workshop on Data Cleaing, Record Linkage, and Object Consolidation, p.40-42, 2003.
66. Rahm E., Do H., Data Cleaning: Problems and Current Approaches, *IEEE Data Eng. Bull.* 23(4): 3-13, 2000.
67. Raman V., Hellerstein J. M., Potter's Wheel: an Interactive Data Cleaning System, *Proc. of Intl. Conf. VLDB*, 2001.
68. Redman T., Data quality: The Field Guide, Digital Press (Elsevier), 2001.
69. Rothenberg J., Metadata to Support Data Quality and Longevity, *Proc. of 1st IEEE Metadata Conf.*, 1996.
70. Santis L.D., Scannapieco M., Catarci T., Trusting Data Quality in Cooperative Information Systems, *Proc. of Intl. Conf. CoopIS,* 2003.
71. Scannapieco M., Pernici B., Pierce E., IP-UML: A Methodology for Quality Improvement based on IP-MAP and UML. Advances in Management Information Systems – Information Quality Monograph (AMIS-IQ), Sharpe M.E., 2004.
72. Schafer J.L., Analysis of Incomplete Multivariate Data, Chapman & Hall, 1997.
73. Schlimmer J., Learning Determinations and Checking Databases, *Proc. of AAAI Workshop on Knowledge Discovery in Databases*, 1991.
74. Sheth A., Wood C., Kashyap V., Q-Data: Using Deductive Database Technology to Improve Data Quality, *Proc. of Intl. Workshop on Programming with Logic Databases (ILPS)*, p. 23-56, 1993.
75. Simitsis A., Vassiliadis P., and Sellis T.K., Optimizing ETL Processes in Data Warehouses, *Proc. of Intl. ICDE Conf.*, 2005
76. Tayi G.K., Ballou D.P., Examining Data Quality, *Com. of the ACM*, 41(2):54-57, 1998.
77. Tan P-N., Kumar V. and Srivastava J., Selecting the Right Interestingness Measure for Association Patterns, *Proc. of Intl. KDD Conf.*, p. 32-41, 2002.
78. Tejada, S., Knoblock, C.A. and Minton S., Learning Object Identification Rules for Information Integration, *Information Systems*, vol. 26, no. 8, 2001.
79. Theodoratos D., Bouzeghoub M., Data Currency Quality Satisfaction in the Design of a Data Warehouse. *Special Issue on Design and Management of Data Warehouses, Int. J. Cooperative Inf. Syst.*, 10(3): 299-326, 2001.
80. Vassiliadis P., Bouzeghoub M., Quix C. Towards Quality-Oriented Data Warehouse Usage and Evolution, *Proc. of Intl. Conf. CAiSE*, p. 164-179, 1999.
81. Vassiliadis P., Vagena Z., Skiadopoulos S., Karayannidis N., ARKTOS: A Tool For Data Cleaning and Transformation in Data Warehouse Environments, *IEEE Data Eng. Bull.,* 23(4): 42-47, 2000.
82. Vassiliadis P., Simitsis A., Georgantas P., Terrovitis M., A Framework for the Design of ETL Scenarios. *Proc. of Intl. Conf. CAiSE*, 2003.
83. Vassiliadis P., Data Warehouse Modeling and Quality Issues, PhD thesis, Technical University of Athens (Greece), 2000.
84. Wang R., Kon H. B., Madnick S. E., Data quality requirements analysis and modeling, Intl. Conf. ICDE, p. 670-677, 1993.
85. Wang R., Storey V., Firth C., A framework for analysis of data quality research, *IEEE TKDE*, 7(4): 670-677, 1995.
86. Wang R., A product perspective on Total Data Quality Management, *Com. of the ACM*, 41(2): 58-65, 1998.

87. Wang R., Journey to Data Quality, vol. 23 of Advances in Database Systems, Kluwer Academic Press, Boston, 2002.
88. Weis M., Naumann F., Detecting Duplicate Objects in XML Documents, *Proc. of the 1st Intl. ACM SIGMOD Workshop on Information Quality in Information Systems (IQIS)*, 2004.
89. Winkler W.E., Data Cleaning Methods, *Proc. of Intl. Conf. KDD*, 2003.
90. Winkler W.E., Methods for Evaluating and Creating Data Quality, *Information Systems*, vol.29, no. 7, 2004.
91. Zaïane O., Han J., Zhu H., Mining Recurrent Items in Multimedia with Progressive Resolution Refinement, *Proc. of IEEE Conf. ICDE*, p.461-476, 2000.
92. Zhu Y., Shasha D., StatStream: Statistical Monitoring of Thousands of Data Streams in Real Time, *Proc. of Intl. Conf. VLDB*, p. 358-369, 2002.