Cost of Low-Quality Data over Association Rules Discovery

Laure Berti-Équille

IRISA
Campus Universitaire de Beaulieu
35042 Rennes Cedex, France
(e-mail: berti@irisa.fr)

Abstract. Quality in data mining critically depends on the preparation and on the quality of processed data sets. Indeed data mining processes and applications require various forms of data preparation (and repair) with several data formatting and cleaning techniques, 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 knowledge. This paper presents a theoretical probabilistic framework for modeling the cost of low-quality data on discovered association rules.

Keywords: Data Quality, Quality of Discovered Association Rules, Minimal Cost Statistical Model.

1 Introduction

In an error-free database or datawarehouse system with perfectly clean data, knowledge discovery techniques (such as clustering, mining association rules or visualization) can be relevantly used from a decisional perspective to automatically derive new knowledge, new concepts, or knowledge patterns from numerical data. Unfortunately, most of the time, these data are neither rigorously chosen from different heterogeneous sources nor carefully controled for quality. Under the general acronym *ETL*, the Extraction-Transformation-Loading activities cover the most prominent tasks of data preparation before the warehousing and mining processes. They include [Vassiliadis *et al.*, 2003]: *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 data set. Data preparation and cleaning processes are complex, costly and critical despite the specialized ETL tools mainly dedicated to relational data available in the market [ETI, 2005], [MS, 2005], [DataMirror, 2005], [ArdentSoftware, 2005]. And the area raised lot of interest with research results [Dasu and Johnson, 2003], [Rahm and Do, 2000], [Winkler, 2003], [Vassiliadis *et al.*, 2003] and several academic tools (Telcordia [Caruso *et al.*, 2000], AJAX [Galhardas *et al.*, 2001],

Potter's Wheel [Raman and Hellerstein, 2001], Arktos [Vassiliadis *et al.*, 2000], IntelliClean [Low *et al.*, 2001], Tailor [Elfeky *et al.*, 2002]).

In the presence of inconsistencies, errors or missing values in the data, it is nevertheless important to estimate the risk of discovering low-quality knowledge by mining low-quality data.

In this paper, our contribution is to present a probabilistic decision model that estimates the cost of discovering low-quality association rules by mining potentially polluted data.

The rest of the paper is organized as follows. Section 2 briefly provides some background information on association rules, data quality and other decision models mainly used in record linkage and data cleaning. Section 3 introduces our decision model and the notation that is used throughout this paper. Section 4 provides concluding remarks and guidelines for future extensions of this work.

2 Background

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 (or transaction) implicates the presence of other distinct set of attributes in the same record (resp. transaction). The quality of association rules is commonly evaluated by looking at their support and confidence. The support of a rule measures the occurence frequency of the pattern in the rule while the confidence is the measure of the strength of implication. Association rule mining is commonly stated as follows: let $I = \{i_1, \ldots, i_n\}$ be a set of *items* and T be a set of data cases. Each data case consists of a subset of items in I. An association rule is an implication of the form $LHS \longrightarrow RHS$, where $LHS \subset I$, $RHS \subset I$, and $LHS \cap RHS = \emptyset$.
The support s of the rule $LHS \longrightarrow RHS$ is measured by the fraction of transactions that contain both LHS and RHS. More formally,

$$s = \frac{\text{number of transactions containing } LHS \cup RHS}{\text{number of transactions}} \tag{1}$$

The confidence c of the rule $LHS \longrightarrow RHS$ states that $c\%$ of transactions that contain LHS also contain RHS and it's the conditional probability of seeing RHS, given that we have seen LHS. More formally,

$$c = \frac{\text{number of transactions containing } LHS \cup RHS}{\text{number of transactions containing } LHS} \tag{2}$$

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, many

other measures for knowledge evaluation have been proposed in the literature with the purpose of supplying subsidies to the user in the understanding and use of the acquired knowledge [Tan *et al.*, 2002], [Lavrac *et al.*, 1999]. Rules may have intrinsic properties (noise tolerance, asymetry, dataset size or dimensionality sensitivity, etc.) or collective properties when considering a set of rules (redundancy, transitivity, consistency, etc.). But, the main drawback of the objective and subjective interestingness measures is to neglect the initial quality of processed data. Data quality is a multidimensional, complex and morphing concept [Dasu and Johnson, 2003]. Table 1 presents some of the dimensions of data quality among more than 200 dimensions that have been proposed in the literature [Wang *et al.*, 1995], [Huang *et al.*, 1999], [Redman, 1996].

Dimension	**Definition**
Availability	Time the data is accessible based on technical equipment and statistics
Freshness	How up-to-date the information is
Accessibility	Estimation of waiting time for information retrieval processing
Security	Estimation of the number of corrupted data
Coverage	Estimation of the number of data for a specific information domain
Accuracy	Estimation of the number of data free-of-error
Completeness	Estimation of the number of missing data or null values
Credibility	User grade based on the reputation of data sources

Table 1. Some Data Quality Dimensions proposed by [Naumann, 2002]

As an illustrative example, one might legitimately wonder whether a so-called "interesting" rule $LHS \longrightarrow RHS$ is meaningful when 30% of the data describing the items of LHS are not up-to-date, 17% of RHS's data are not accurate, 14% of LHS's data come from sources that have bad credibility. In this paper, we consider that identifying interesting rules should also take into account the quality of underlying data used by the rule mining process: despite high interestingness measures, there are interesting rules discovered from dirty data, others from clean data, but they don't have the same added-value. This can be seen as a classification problem where the goal is to correctly assign cases (measurements, observations, etc.) to one of a finite number of classes. Most of the currently available algorithms for classification are designed to minimize error rate, *i.e.*, the number of incorrect predictions made. This implicity assumes that all errors are equally costly. In our context, there are many different types of cost involved on the selection of discovered rules. For instance, discovering interesting rules from inaccurate data may not have the same cost (or impact) than discovering rules from out-of-date data. In this study, we consider only the cost of misclassification error which is related to assigning different weights to different

misclassification errors. Misclassification costs may be generally described by an arbitrary cost matrix C, with elements of the form c_{ij}, meaning the cost of predicting that an example belongs to the class i when in fact it belongs to j. The Bayesian decision approach is based on the assumption that the decision problem is posed in probabilistic terms, and that all the relevant probability values are known. In this paper, we propose a constant error cost Bayesian model which means that the cost of a certain type of error may be constant. In some cases, we are uncertain about the actual costs. To account for this uncertainty, we can use a probability distribution over a range of possible costs. To keep the presentation simple, we do not consider probability distributions over costs in this study. Our work is correlated to several works in data cleaning and Table 2 presents several decision models proposed in the literature mainly for record linkage. Our model is similar to the one proposed by Verykios *et al.* [Verykios *et al.*, 2003] as it minimizes the cost of making a decision rather than the probability of error in a decision of record matching. Our contribution is to adapt this model for association rule mining and for minimizing the cost of the rule selection in presence of low-quality data and of a misclassification region that can occur when erroneous data can be classified correct because they're in the range of correct values and correct data can be classified erroneous because they're in the range of erroneous values or outliers.

Model *(Tool)*	**Authors**	**Type of Model**
Error-based Model	[Fellegi and Sunter, 1969]	Probabilistic
EM-based Method	[Dempster *et al.*, 1977]	Probabilistic
Bayesian Cost-based Model	[Verykios *et al.*, 2003]	Probabilistic
Induction	[Bilenko and Mooney, 2003]	Probabilistic
Clustering for Record Linkage *(Tailor)*	[Elfeky *et al.*, 2002]	Probabilistic
1-1 matching	[Winkler, 2004]	Probabilistic
Bridging File	[Winkler, 2003]	Probabilistic
sorted-NN method	[Hernandez and Stolfo, 1995]	Empirical
XML Object Matching	[Weis and Naumann, 2004]	Empirical
Hierarchical Structure *(Delphi)*	[Ananthakrishna *et al.*, 2002]	Empirical
Matching Prediction based on clues	[Buechi *et al.*, 2003]	Knowledge-based
Functional Dependencies Inference	[Lim *et al.*, 1993]	Knowledge-based
Transformation functions *(Active Atlas)*	[Tejadaa *et al.*, 2001]	Knowledge-based
Rules and sorted-NN *(Intelliclean)*	[Low *et al.*, 2001]	Knowledge-based

Table 2. Decision Models for Record Linkage and Duplicate Identification

3 Cost-based Probabilistic Model

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

3.1 Definition and Notations

Consider the item $x \in LHS \cup RHS$ of a given rule, we use $P_{CE}(x)$ to denote the probability that the item x will be classified as "erroneous" (or "polluted") *wrt* to one or more quality dimensions relevant to the application, 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 dimensions). 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). Intuitively, the item x can be an attribute whose quality dimensions are measured and aggregated from all the existing values of the attribute domain.

For an arbitrary average quality vector $\bar{q} \in Q$ on all data items in LHS $\cup RHS$ of the rule, we denote by $P(\bar{q} \in Q|CC)$ or $f_{CC}(\bar{q})$ the conditional probability of the pattern $\bar{q}$ that corresponds to the average of quality vectors of the items that are classified as correct (CC). Similarly, we denote by $P(\bar{q} \in Q|CE)$ or $f_{CE}(\bar{q})$ the conditional probability of the pattern $\bar{q}$ corresponds to the average of quality vectors of the items that are classified erroneous (CE). We denote by d the decision of the predicted class of the rule (*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 $P(d = D_i|s = j)$ correspondingly, the joint and the conditional probability that the decision D_i is taken, when the actual status of data quality (*CC, CE, AE, AC*) is j. We also denote by c_{ij} the cost of making a decision D_i for classifying a rule with actual data quality status j of the items sets composing the parts of the rule.

3.2 Cost-based Bayesian Decision Model

Based on the example in Table 3 where we can see how the cost of different decisions could affect the result of selection among interesting rules, we need to minimize the mean cost $\bar{c}$ 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 3. Costs of various decisions for classifying interesting rules

The mean cost is written as follows:

$$\begin{aligned}\bar{\mathrm{c}}=&\mathrm{c}_{10}.P(d=D_1,s=CC)+c_{20}.P(d=D_2,s=CC)+c_{30}.P(d=D_3,s=CC)\\&+c_{11}.P(d=D_1,s=CE)+c_{21}.P(d=D_2,s=CE)+c_{31}.P(d=D_3,s=CE)\\&+c_{12}.P(d=D_1,s=AE)+c_{22}.P(d=D_2,s=AE)+c_{32}.P(d=D_3,s=AE)\\&+c_{13}.P(d=D_1,s=AC)+c_{23}.P(d=D_2,s=AC)+c_{33}.P(d=D_3,s=AC)\end{aligned} \tag{3}$$

From the Bayes theorem, the following is true:

$$P(d=D_i,s=j)=P(d=D_i|s=j).P(s=j) \tag{4}$$

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

$$P(d=D_i|s=j)=\sum_{\bar{q}\in D_i} f_j(\bar{q}) \tag{5}$$

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 else $P(s=CC)$ 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 misclassification region $P(s=CE)$ could be simplified as $1-\pi^0$.

The mean cost $\bar{\mathrm{c}}$ in Eq. 3 based on Eq. 4 is written as follows:

$$\begin{aligned}\bar{c} = & c_{10}.P(d = D_1|s = CC).P(s = CC) \\ & + c_{20}.P(d = D_2|s = CC).P(s = CC) + c_{30}.P(d = D_3|s = CC).P(s = CC) \\ & + c_{11}.P(d = D_1|s = CE).P(s = CE) \\ & + c_{21}.P(d = D_2|s = CE).P(s = CE) + c_{31}.P(d = D_3|s = CE).P(s = CE) \\ & + c_{12}.P(d = D_1|s = AE).P(s = AE) \\ & + c_{22}.P(d = D_2|s = AE).P(s = AE) + c_{32}.P(d = D_3|s = AE).P(s = AE) \\ & + c_{13}.P(d = D_1|s = AC).P(s = AC) \\ & + c_{23}.P(d = D_2|s = AC).P(s = AC) + c_{33}.P(d = D_3|s = AC).P(s = AC)\end{aligned} \tag{6}$$

and by using Eq. 5 and by dropping the dependent vector variable $\bar{q}$, Eq. 6 becomes:

$$\begin{aligned}\bar{c} = & \sum_{\bar{q}\in D_1}[f_{CC}.c_{10}.\pi^0 + f_{CE}.c_{11}.(1 - \pi^0 - \pi^0_{AC} + \pi^0_{AE}) \\ & + f_{AE}.c_{12}.\pi^0_{AE} + f_{AC}.c_{13}.\pi^0_{AC}] \\ & + \sum_{\bar{q}\in D_2}[f_{CC}.c_{20}.\pi^0 + f_{CE}.c_{21}.(1 - \pi^0 - \pi^0_{AC} + \pi^0_{AE}) \\ & + f_{AE}.c_{22}.\pi^0_{AE} + f_{AC}.c_{23}.\pi^0_{AC}] \\ & + \sum_{\bar{q}\in D_3}[f_{CC}.c_{30}.\pi^0 + f_{CE}.c_{31}.(1 - \pi^0 - \pi^0_{AC} + \pi^0_{AE}) \\ & + f_{AE}.c_{32}.\pi^0_{AE} + f_{AC}.c_{33}.\pi^0_{AC}]\end{aligned} \tag{7}$$

Every point $\bar{q}$ in the decision space D, belongs either in partition 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 $\bar{q}$ is assigned to the three optimal areas as follows:

To D^0_1 if:
$f_{CC}.c_{10}.\pi^0 + f_{CE}.c_{11}.(1 - \pi^0 - \pi^0_{AC} + \pi^0_{AE}) + f_{AE}.c_{12}.\pi^0_{AE} + f_{AC}.c_{13}.\pi^0_{AC}$
$\leq f_{CC}.c_{30}.\pi^0 + f_{CE}.c_{31}.(1 - \pi^0 - \pi^0_{AC} + \pi^0_{AE}) + f_{AE}.c_{32}.\pi^0_{AE} + f_{AC}.c_{33}.\pi^0_{AC}$
and,
$f_{CC}.c_{10}.\pi^0 + f_{CE}.c_{11}.(1 - \pi^0 - \pi^0_{AC} + \pi^0_{AE}) + f_{AE}.c_{12}.\pi^0_{AE} + f_{AC}.c_{13}.\pi^0_{AC}$
$\leq f_{CC}.c_{20}.\pi^0 + f_{CE}.c_{21}.(1 - \pi^0 - \pi^0_{AC} + \pi^0_{AE}) + f_{AE}.c_{22}.\pi^0_{AE} + f_{AC}.c_{23}.\pi^0_{AC}$.

To D^0_2 if:
$f_{CC}.c_{20}.\pi^0 + f_{CE}.c_{21}.(1 - \pi^0 - \pi^0_{AC} + \pi^0_{AE}) + f_{AE}.c_{22}.\pi^0_{AE} + f_{AC}.c_{23}.\pi^0_{AC}$
$\leq f_{CC}.c_{30}.\pi^0 + f_{CE}.c_{31}.(1 - \pi^0 - \pi^0_{AC} + \pi^0_{AE}) + f_{AE}.c_{32}.\pi^0_{AE} + f_{AC}.c_{33}.\pi^0_{AC}$
and,
$f_{CC}.c_{20}.\pi^0 + f_{CE}.c_{21}.(1 - \pi^0 - \pi^0_{AC} + \pi^0_{AE}) + f_{AE}.c_{22}.\pi^0_{AE} + f_{AC}.c_{23}.\pi^0_{AC}$
$\leq f_{CC}.c_{10}.\pi^0 + f_{CE}.c_{11}.(1 - \pi^0 - \pi^0_{AC} + \pi^0_{AE}) + f_{AE}.c_{11}.\pi^0_{AE} + f_{AC}.c_{13}.\pi^0_{AC}$.

To D^0_3 if:
$f_{CC}.c_{30}.\pi^0 + f_{CE}.c_{31}.(1 - \pi^0 - \pi^0_{AC} + \pi^0_{AE}) + f_{AE}.c_{32}.\pi^0_{AE} + f_{AC}.c_{33}.\pi^0_{AC}$
$\leq f_{CC}.c_{10}.\pi^0 + f_{CE}.c_{11}.(1 - \pi^0 - \pi^0_{AC} + \pi^0_{AE}) + f_{AE}.c_{12}.\pi^0_{AE} + f_{AC}.c_{13}.\pi^0_{AC}$

and,
$f_{CC}.c_{30}.\pi^0 + f_{CE}.c_{31}.(1-\pi^0-\pi^0_{AC}+\pi^0_{AE}) + f_{AE}.c_{32}.\pi^0_{AE} + f_{AC}.c_{33}.\pi^0_{AC}$
$\leq f_{CC}.c_{20}.\pi^0 + f_{CE}.c_{21}.(1-\pi^0-\pi^0_{AC}+\pi^0_{AE}) + f_{AE}.c_{22}.\pi^0_{AE} + f_{AC}.c_{23}.\pi^0_{AC}.$

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:

$$D_1^0 = \{\bar{q} : \frac{f_{CE}}{f_{CC}} \leq \frac{\pi^0}{1-\pi^0}.\frac{c_{30}-c_{10}}{c_{11}-c_{31}} \text{ and, } \frac{f_{CE}}{f_{CC}} \leq \frac{\pi^0}{1-\pi^0}.\frac{c_{20}-c_{10}}{c_{11}-c_{21}}\} \quad (8)$$

$$D_2^0 = \{\bar{q} : \frac{f_{CE}}{f_{CC}} \geq \frac{\pi^0}{1-\pi^0}.\frac{c_{20}-c_{10}}{c_{11}-c_{21}} \text{ and, } \frac{f_{CE}}{f_{CC}} \leq \frac{\pi^0}{1-\pi^0}.\frac{c_{30}-c_{20}}{c_{21}-c_{31}}\} \quad (9)$$

$$D_3^0 = \{\bar{q} : \frac{f_{CE}}{f_{CC}} \geq \frac{\pi^0}{1-\pi^0}.\frac{c_{30}-c_{10}}{c_{11}-c_{31}} \text{ and, } \frac{f_{CE}}{f_{CC}} \geq \frac{\pi^0}{1-\pi^0}.\frac{c_{30}-c_{20}}{c_{21}-c_{31}}\} \quad (10)$$

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

$$L = \frac{\pi^0}{1-\pi^0}.\frac{c_{30}-c_{10}}{c_{11}-c_{31}} \quad (11)$$

$$P = \frac{\pi^0}{1-\pi^0}.\frac{c_{20}-c_{10}}{c_{11}-c_{21}} \quad (12)$$

$$N = \frac{\pi^0}{1-\pi^0}.\frac{c_{30}-c_{10}}{c_{11}-c_{31}} \quad (13)$$

3.3 Minimal and Maximal Quality of Association Rules for Correctly Classified Data

For categorical quality dimensions, errors or pollutions (*non-quality*) in the data sets (e.g., in LHS and RHS parts of the rules) might be measured using contingency table approach where the item subsets with actual and estimated quality for a selected sample of items. It is then possible to calculate the proportion of data items that are correctly classified and estimated the quality of the entire set based on inferential statistics. Another more sophisticated approach can utilizes a random-stratified sampling method in which the same number of samples is chosen from each item subsets. This has the advantage that minor item subsets are not under-represented in the sample, which makes it possible to calculate the average quality of individual item sets.

We present now a model in which the quality of a given rule r, $P_{CC}[\bar{Q}_r]$ is defined as the probability of data items in the left and right-hand sides data sets of the rule that are correctly classified. Given two data sets with average qualities of $P_{CC}[\bar{Q}_{LHS}]$ and $P_{CC}[\bar{Q}_{RHS}]$, the quality of the rule $P_{CC}[\bar{Q}_r]$, is given by:

$$P_{CC}[\bar{Q}_r] = P_{CC}[\bar{Q}_{LHS}].P_{CC}[\bar{Q}_{RHS}|\bar{Q}_{LHS}] \tag{14}$$

The conditional probability $P_{CC}[\bar{Q}_{RHS}|\bar{Q}_{LHS}]$ is the probability of correctly classified data items in LHS that are also correclty classified in RHS. The equation can be expanded for situations involving more than two item sets composing the rule.

From the preceding equations, the maximum and minimum quality of a given association rule can be determined based on the average quality of the several item sets I_i composing the rule.
Maximum quality is given by:

$$P_{CC}[\bar{Q}_r^{max}] = min\{P[\bar{Q}_{I_i}]\} \text{ with } i = 1, 2, \ldots, n \tag{15}$$

Minimum quality is given by:

$$P_{CC}[\bar{Q}_r^{min}] = max\{0, (1 - \sum_{i=1}^{n} P_{CE}[\bar{Q}_{I_i}])\} \tag{16}$$

where $P_{CE}[\bar{Q}_{I_i}]$ is the average quality probability of the items in the data set I_i that are classified erroneous. These formulae lead to several general conclusions about composite rule quality. Composite rule quality will at the best be equal to the quality of the least quality data set. At worst composite rule quality will be equal to one minus the sum of the probability of misclassified items on each data set (or to zero if this value is negative).

4 Conclusion

This paper presents a prospective work on a theoretical probabilistic framework for estimating the cost of low-quality data on discovered association rules. Our future plans regarding this work, are to study the optimality of our decision model, to propose error estimation and to validate the model with experiments on large data sets and discovered rules with several multi-dimensional quality metrics.

References

[Ananthakrishna *et al.*, 2002]R. Ananthakrishna, S. Chaudhuri, and V. Ganti. Eliminating fuzzy duplicates in data warehouses. In *Proc. of the 28th Intl. Conf. on Very Large Data Bases (*VLDB*)*, Hong Kong, China, 2002.

[ArdentSoftware, 2005]Datastage suite. Available at http://www.ardentsoftware.com/, 2005.

[Bilenko and Mooney, 2003]Mikhail Bilenko and Raymond J. Mooney. Adaptive duplicate detection using learnable string similarity measures. In *KDD*, pages 39–48, 2003.

[Buechi *et al.*, 2003]M. Buechi, A. Borthwick, A. Winkel, and A. Goldberg. Cluemaker: a language for approximate record matching. In *Proc. of the 8th Inl. Conf. on Information Quality (*ICIQ *2003)*, Boston, MA, 2003.

[Caruso *et al.*, 2000]F. Caruso, M. Cochinwala, U. Ganapathy, G. Lalk, and P. Missier. TELCORDIA's database reconciliation and data quality analysis tool. In *Proc. of the Intl. Conf. on Very Large Data Bases (*VLDB*)*, 2000.

[Dasu and Johnson, 2003]T. Dasu and T. Johnson. *Exploratory Data Mining and Data Cleaning*. Wiley, 2003.

[DataMirror, 2005]Transformation server. Available at http://www.datamirror.com/, 2005.

[Dempster *et al.*, 1977]A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *ournal of the Royal Statistical Society*, 39:1–38, 1977.

[Elfeky *et al.*, 2002]M.G. Elfeky, V.S. Verykios, and A.K. Elmagarmid. Tailor: A record linkage toolbox. In *Proc. of the Intl. Conf. on Data Engineering (*ICDE*)*, 2002.

[ETI, 2005]ETI*EXTRACT. Available at http://www.eti.com/, 2005.

[Fellegi and Sunter, 1969]I.P. Fellegi and A.B. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64, 1969.

[Galhardas *et al.*, 2001]H. Galhardas, D. Florescu, D. Shasha, E. Simon, and C. Saita. Declarative data cleaning: Language, model and algorithms. In *Proc. of the Intl. Conf. on Very Large Data Bases (*VLDB*)*, pages 371–380, 2001.

[Hernandez and Stolfo, 1995]M. Hernandez and S. Stolfo. The merge/purge problem for large databases. In *Proc. of* ACM *Special Interest Group on Management of Data Intl. Conf. (*SIGMOD *1995)*, San Jose, California, 1995.

[Huang *et al.*, 1999]K. Huang, Y. Lee, and R. Wang. *Quality Information and Knowledge Management*. Prentice Hall, New Jersey, 1999.

[Lavrac *et al.*, 1999]Nada Lavrac, Peter A. Flach, and Blaz Zupan. Rule evaluation measures: A unifying view. In *ILP*, pages 174–185, 1999.

[Lim *et al.*, 1993]L. Lim, J. Srivastava, S. Prabhakar, and J. Richardson. Entity identification in database integration. In *Proc. of the Intl. Conf. on Data Engineering (*ICDE*)*, Wien, Austria, 1993.

[Low *et al.*, 2001]W.L. Low, M.L. Lee, and T.W. Ling. A knowledge-based approach for duplicate elimination in data cleaning. *Information System*, 26(8), 2001.

[MS, 2005]MS data transformation services. Available at http://www.microsoft.com/sql/evaluation/features/datatran.asp, 2005.

[Naumann, 2002]F. Naumann. *Quality-Driven Query Answering for Integrated Information Systems.*, volume 2261 of *Lecture Notes in Computer Science*. Springer-Verlag, 2002.

[Rahm and Do, 2000]E. Rahm and H. Do. Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, 23(4):3–13, 2000.

[Raman and Hellerstein, 2001]V. Raman and J. M. Hellerstein. Potter's wheel: an interactive data cleaning system. In *Proc. of the Intl. Conf. on Very Large Data Bases (*VLDB*)*, 2001.

[Redman, 1996]T.C. Redman. *Data Quality for the Information Age.* Artech House, 1996. ISBN 0-89006-8836.

[Tan *et al.*, 2002]Pang-Ning Tan, Vipin Kumar, and Jaideep Srivastava. Selecting the right interestingness measure for association patterns. In *KDD*, pages 32–41, 2002.

[Tejadaa *et al.*, 2001]S. Tejadaa, C.A. Knoblock, and S. Minton. Learning object identification rules for information integration. *Information Systems*, 26(8), 2001.

[Vassiliadis *et al.*, 2000]P. Vassiliadis, Z. Vagena, S. Skiadopoulos, and N. Karayannidis. ARKTOS: A tool for data cleaning and transformation in data warehouse environments. *IEEE Data Eng. Bull.*, 23(4):42–47, 2000.

[Vassiliadis *et al.*, 2003]P. Vassiliadis, A. Simitsis, P. Georgantas, and M. Terrovitis. A framework for the design of ETL scenarios. In *Proc. of the 15th Conf. on Advanced Information Systems Engineering (*CAiSE*'03)*, Klagenfurt, Austria, 2003.

[Verykios *et al.*, 2003]V.S. Verykios, G.V. Moustakides, and M.G. Elfeky. A bayesian decision model for cost optimal record matching. *VLBD Journal*, 12(4):28–40, 2003.

[Wang *et al.*, 1995]R. Wang, V. Storey, and C. Firth. A framework for analysis of data quality research. IEEE *Transactions on Knowledge and Data Engineering*, 7(4):623–638, 1995.

[Weis and Naumann, 2004]M. Weis and F. Naumann. Detecting duplicate objects in XML documents. In *Proc. of the 1st Intl. ACM SIGMOD Workshop on Information Quality in Information Systems*, 2004.

[Winkler, 2003]W.E. Winkler. Data cleaning methods. In *Proc. of the Intl. Conf. KDD*, 2003.

[Winkler, 2004]W.E. Winkler. Methods for evaluating and creating data quality. *Information Systems*, 29(7), 2004.