AN ANALYTICAL FRAMEWORK TO ANALYZE DEPENDENCIES AMONG DATA QUALITY DIMENSIONS

(Complete - paper)

Fabrizio De Amicis
BSI Bank, Switzerland
fabrizio.deamicis@bsi.ch

Daniele Barone
Università di Milano “Bicocca”, Italy
Dipartimento di Informatica, Sistemistica e Comunicazione
barone@disco.unimib.it

Carlo Batini
Università di Milano “Bicocca”, Italy
Dipartimento di Informatica, Sistemistica e Comunicazione
batini@disco.unimib.it

Abstract: Dependencies among data quality dimensions, although they are not thoroughly and deeply analyzed in the research area of information quality, represent an issue of primary importance. The knowledge on dependencies can help the data quality analyst in finding errors, it can be useful to assess the quality level of a data set, which can improve knowledge on data quality dimensions. This paper proposes a data-driven approach for the analysis of dependencies among data quality dimensions. The analytical framework proposed here provides the main models of dependencies and analytic formulae, based on the entropy of Shannon. It characterizes each model, and proposes measures of correlation among the dimensions. Automatic identification of models of dependencies for a given data set is performed with the implemented version of the framework. Examples and case study are provided.

Key words: Dependencies among data quality dimensions, entropic correlation measure, data quality meta-model, data quality analysis

1 INTRODUCTION

During the last years, one of the main issues of the research in information systems and data bases has been the quality of the data used in business, operational and decisional business processes. Several methodologies and guidelines have been proposed to address this issue. The main general-purpose methodologies are the TDQM and TIQM methodologies (see [7], [8], [25], [28], [29]). The core concept of TDQM is to consider data as a particular output of manufacturing processes. TDQM proposes four phases to manage a generic information product (IP) along its life cycle: i) model the IP and choose the quality dimensions, ii) measure actual values of dimensions, iii) analyze the causes of errors and the IP production process, and iv) identify a DQ improvement process. The main contributions of TIQM are a set of specific techniques for cost-benefit analysis and a general managerial perspective. The cost-benefit analysis is conducted by analyzing non-quality information costs, finding main causes of non quality, and conceiving projects to improve quality. However, both in TDQM and in TIQM measurement and improvement activities are performed on a set of quality dimensions (e.g. accuracy, completeness, and
timeliness), which are measured and analyzed independently of each other. Frequently in data management, several correlations exist among quality dimensions. E.g. in a web site, if we want to public complete data, we have to sacrifice timeliness. In this paper we investigate the issue of dependencies among dimensions. In probability theory, two events are considered independent if the occurrence of one event makes it neither more nor less probable that the other occurs. Formally, two event A and B are independent if \( P(AB) = P(A)P(B) \), where \( P(AB) \) is the probability that both events A and B occur. Intuitively, we say that a dependency exists between two dimensions \( d_1 \) and \( d_2 \) if there is some degree of relationship among values of dimension \( d_1 \), and values of dimension \( d_2 \). This paper uses the concept and the properties of entropy to analyze the dependencies among dimensions. The main models of dependencies: i) perfect dependency, ii) independency, and iii) partial dependency are presented and characterized via analytical formulas. Considering the dependencies among data quality dimensions, besides their values judged independently, leads to several potential advantages in the measurement and improvement process: i) it chooses the most effective improvement activity on involved dimensions, unfolding progressively the effects from independent to dependent dimensions, and ii) it minimizes the cost of the improvement process, since redundant and overlapping activities are avoided. Moreover, dependencies may contribute to discover cause-effect patterns on data quality dimensions, and to find the most relevant input-source of errors in the data.

The paper represents one of the first approaches for the data-driven analysis of dependencies among data quality dimensions. For this goal, we propose an analytical framework, based on the entropy of Shannon, and statistical measures of correlation among dimensions. In the following, we assume the relational model of data as data model, in which attributes are grouped in tables. Figure 1 shows the main elements of the analytical framework at a high level. Starting from a generic data set, we apply data quality assessment activities (see [3], [7], [6], [15], [19], and [21] for assessment methodologies proposed in the literature) to define a model for the quality of attribute values (data quality meta-model). The quality model is composed of meta-attributes (see [27]) that represent the quality of the attribute values, i.e. the quality related to each observation of the attributes. Exploiting the meta-model, we use the analytical framework to analyze and to extract knowledge on the dependencies among dimensions. In this paper the assessment activity is performed using existing methodologies, while the usage of dependencies in the improvement process is out of the scope of this paper. The analysis of dependencies can be performed: i) on the same attribute, ii) on different attributes of the same table, iii) on attributes belonged to different tables.

![Figure 1 Main elements of the analytical framework](image)

The paper is organized as follows. Section 2 presents the related work. Section 3 provides definitions of the quality meta-model and introduces basic definitions of entropy. Section 4 presents the first step of the framework analysis. Section 5 presents the main dependencies models. A SAS® program has been developed by the authors for automatic identification of dependencies models. In particular we have applied the SAS® program to Moody’s and Standard & Poor’s bond rating variables, stored in a specific internal financial database. The analysis and results of experiment are presented in Section 6. Section 7 presents conclusions and future work.
2 RELATED WORK
The analysis of dependencies in the area of data quality has been investigated so far in several works, mainly in terms of tradeoffs among dimensions. An example is illustrated in [2], where the accuracy-timeliness tradeoff is investigated in terms of how the improvement of timeliness could adversely affect the accuracy. In the context of decision making and review, a theoretical framework is proposed for the analysis of tradeoffs, with the goal of determining appropriate changes to data managed in information systems. An accuracy-timeliness utility function figures prominently in the analysis. When it is not possible to determine the utility function, the authors propose an approach to identify a suitable approximation. Another type of tradeoff is proposed in [14], among timeliness, accuracy and cost. The aim is to find heuristics to balance these three factors, through a middleware framework for applications requiring real-time information collection and exchange in distributed real-time environments. A practical case of tradeoff among timeliness and other generic data quality dimensions is analyzed in [23], in the context of the Bureau of Labor Statistics Covered Employment and Wages. The objective is to determine if data quality decreases as a result of receiving data earlier than the current due dates. No significant quality deterioration is shown.

A framework that allows the systematic exploration of tradeoff between completeness and consistency is presented in [3]. The relative weight (importance) of completeness and consistency to the decision maker is an input to the analysis. In order to examine the tradeoff, the authors explore various facets of the two dimensions, that produce analytical expressions for the measurement activity. The utility of various combinations of completeness and consistency for fixed and variable budgets provides guidance to evaluate the appropriate tradeoffs of these factors for specific decision contexts. For this goal, a utility function, similar to that presented in [2], produces an optimized value for tradeoff. In the theoretical frameworks presented in [2] and [3], the tradeoff is often based on i) a weight, that represents the importance of a dimension versus another dimension, and on ii) a functional relationship, that defines the bind between the dimensions involved. In order to analyze the tradeoff, the framework requires both types of information; furthermore, they have to be provided by the user, and are not calculated automatically by the framework itself. When the functional relationships are known or can be determined, mathematical models can be used to balance the tradeoff. In case the functional relationship is not available, the framework suggests an alternative approach based on the use of generic families of functions to approximate the required real function. Even in this case functional parameters must be provided by the user.

In [15] a rigorous and pragmatic methodology for information quality assessment, called AIMQ, is presented. The AIMQ methodology consists of three components: i) a 2×2 model or framework of what information quality means to information consumers and managers; ii) a questionnaire for measuring information quality along the dimensions; iii) analysis techniques for interpreting the assessments gathered by the questionnaire. The analysis results of the methodology highlight that information quality is a single phenomenon where dimensions are not inherently independent. The table of correlations among dimensions, calculated on the basis of answers, is reported in the paper. In [12], within the contexts of a business situation, a set of logical interdependencies among dimensions, is presented. The authors define a hierarchical result-oriented taxonomy of data quality dimensions (also called attributes) composed of direct and indirect attributes. Direct attributes represent the main fundamental dimensions and, when their values change, they directly influence the results of business operations. The indirect attributes determine/contribute to the direct attributes; hence indirectly influence the results. An example of a direct attribute is the significantly relevant attribute, that measures the degree of relevance of an individual data/information items in a specific decision situation. An example of an indirect attribute, related to the significantly relevant attribute, is the currency attribute. Indeed, when information is not current from the point of view of the decision situation, then it is irrelevant.
As previously illustrated, the correlation among data quality dimensions has been investigated in the literature focusing on tradeoff analysis, methodology assessment and logical interdependence. Our analytical framework uses a data-driven approach to detect the dependencies among dimensions directly from data. For this goal, we do not use any a priori-knowledge except the knowledge resulting from the assessment activity, stored in the quality meta-model. In the formulation of the analytical framework, methodologies for the analysis of categorical data have been largely considered (see [1], [10], [11], [3], [9]). To identify and measure dependencies among categorical data, a large number of correlation measures and statistical tests have been introduced in statistics and information theory (see [1], [5], [17], [13], [20]). Here we propose an analytical framework composed by models of dependencies where each model is characterized by an analytical entropic equation. Our approach proposes innovative measures of entropic values, as the identification of the most occurring data quality dimension via entropic value, presented in Section 4, and techniques to detect models of dependencies, as the collapsing techniques used to identify the partial synonymy model and the lower-resolution absolute synonymy model presented in Section 5.

3 DEFINITION OF DATA QUALITY META-MODEL

This section defines the data quality meta-model used to analyze dependencies among dimensions; for such a goal we first introduce some basic definitions of entropy. The meta-model is applied to a generic data set, made of a set of N observations (1,2,...,i,...,N), each one made in turn of \(v\) attribute values \((X_1,Y_1,...,Z_i)\). The data quality dimensions considered in the meta-model are taken from a set of \(\{d_1,d_2,...,d_m\}\). E.g. referring to [6], we may have \(D = \{\text{syntactic accuracy}, \text{semantic accuracy}, \text{completeness},..., \text{uniqueness} \}\). Our meta-model is organized in terms of a set of variables, one for each attribute in the data set, that contain the dimensions name of the dimensions detected in the assessment of the attributes. More formally, let \(X\) be a generic attribute; we introduce in the meta-model a variable \(D^X\), associated to \(X\), called "Data Quality Dimension Variable (DQDV)" for each observation \(i\) defined as follows.

\[
X^D_i = \begin{cases} 
- \text{Not_error_detected if no dimension in the set } d_1,d_2,...,d_m \\
\text{is characterized by an error in the observation } i \\
- \text{the list of dimensions characterized by an error in the observation } i, \text{ otherwise}
\end{cases}
\]

For simplicity, in the following examples for each observation we make the assumption that the dimensions affected by an error is at most one.

Table 1 provides an example of the above definition, in which observations are referred to a unique attribute “Bond rating”, that represents the financial risk related to bond instrument. We assume that a data quality assessment is performed \textit{a priori} in order to feed the \(X^D\) values.

<table>
<thead>
<tr>
<th>Observation</th>
<th>(X) (bond rating)</th>
<th>(X^D) (DQDV of the bond rating)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AAA</td>
<td>Timeliness</td>
</tr>
<tr>
<td>2</td>
<td>AA</td>
<td>Not_error_detected</td>
</tr>
<tr>
<td>3</td>
<td>NULL</td>
<td>Completeness</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td>...</td>
</tr>
</tbody>
</table>

Table 1. Example of \(X^D\) values for the attribute bond rating
We now introduce Shannon’s entropy definitions (see [1], [26]), particularly suitable for the analysis of data dependencies, with the final goal to define the concept of data quality dimension dependency between two attributes X and Y, for the set of dimensions \{d_1, d_2, \ldots, d_m\}.

**Definition 1.** The entropy associated to a variable \(X^D\) is the measure of information and uncertainty of the variable \(X^D\). Formally, the entropy is defined by the Shannon formula

\[
H_{X^D}(p) = -\sum_{j=1}^{m} P(X^D = d_j) \log P(X^D = d_j)
\]

where \(p\) is the vector of probabilities \(P(X^D = d_j) = p_j\) for the dimensions \(\{d_1, d_2, \ldots, d_m\}\), that represent the set of values that \(X^D\) can take.

**Definition 2.** The entropy between \(X^D\) and \(Y^D\) is defined as

\[
H_{X^D Y^D}(p) = -\sum_{j=1}^{m} \sum_{k=1}^{m} P(X^D = d_j, Y^D = d_k) \log P(X^D = d_j, Y^D = d_k)
\]

where \(p\) is the vector of the joint probabilities \(P(X^D = d_j, Y^D = d_k) = p_{jk}\) for the dimensions \(\{d_1, d_2, \ldots, d_m\}\).

**Definition 3.** The conditional entropy of \(X^D\) to \(Y^D = d_k\) is defined by the following formula

\[
H(X^D|Y^D = d_k) = \sum_{j=1}^{m} P(X^D = d_j|Y^D = d_k) \log \left( \frac{1}{P(X^D = d_j|Y^D = d_k)} \right)
\]

**Definition 4.** The average mutual information, a measure of association between two variables, is defined as follow:

\[
I_{X^D Y^D}(p) = -\sum_{j=1}^{m} \sum_{k=1}^{m} P(X^D = d_j, Y^D = d_k) \log \left( \frac{P(X^D = d_j, Y^D = d_k)}{P(X^D = d_j)P(Y^D = d_k)} \right)
\]

Sections 4 and 5 present techniques and models that use definitions 1 to 4 to formally express the correlation between errors in X and the corresponding errors in Y.

### 4 Preliminary Analysis of the Dependencies

This section presents the first step of the framework analysis and proposes a method to characterize the frequencies of occurrences of errors (in a set of attributes X, Y,\ldots, Z), via entropic values. Given a variable \(X^D\), the number of observations of a generic dimension \(d_j\) is:

\[
N(d_j) = \sum_{i=1}^{N} F_{[d_j]}(X^D_i)
\]

where N is the total number of observations of the variable \(X^D\), and

\[
F_{[d_j]}(X^D_i) = \begin{cases} 1, & \text{if } X^D_i = d_j \\ 0, & \text{if } X^D_i \neq d_j \end{cases}
\]

More general, in the case of a set of attribute X, Y,\ldots, Z, the frequency of a generic combination \((d_1, d_2, \ldots, d_i)\) related to \(X^D, Y^D, \ldots, Z^D\) is
\[
N(d_j,d_k,...,d_l) = \sum_{i=1}^{N} F_{d_j,d_k,...,d_l}\left(X_i^D, Y_i^D, ..., Z_i^D\right)
\]

where
\[
F_{d_j,d_k,...,d_l}\left(X_i^D, Y_i^D, ..., Z_i^D\right) = \begin{cases}
1, & \text{if } \left(X_i^D, Y_i^D, ..., Z_i^D\right) = \{d_j, d_k, ..., d_l\} \\
0, & \text{if } \left(X_i^D, Y_i^D, ..., Z_i^D\right) \neq \{d_j, d_k, ..., d_l\}
\end{cases}
\]

An example of the preliminary analysis output, related to two variables (Moody’s and Standard & Poor’s bond rating) of a financial database, is presented in Table 2.

<table>
<thead>
<tr>
<th>(X^D) related to attribute rating Moody’s</th>
<th>(Y^D) related to attribute rating S&amp;P</th>
<th>Count</th>
<th>Percent</th>
<th>Entropic Value</th>
<th>Normalized entropic Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not_error_detected</td>
<td>Not_error_detected</td>
<td>2937</td>
<td>70.35</td>
<td>0.608</td>
<td>0.877</td>
</tr>
<tr>
<td>Not_error_detected</td>
<td>Timeliness</td>
<td>518</td>
<td>12.41</td>
<td>0.375</td>
<td>0.541</td>
</tr>
<tr>
<td>Not_error_detected</td>
<td>Completeness</td>
<td>252</td>
<td>6.04</td>
<td>0.228</td>
<td>0.329</td>
</tr>
<tr>
<td>Timeliness</td>
<td>Not_error_detected</td>
<td>223</td>
<td>5.34</td>
<td>0.208</td>
<td>0.301</td>
</tr>
<tr>
<td>Timeliness</td>
<td>Timeliness</td>
<td>132</td>
<td>3.16</td>
<td>0.140</td>
<td>0.202</td>
</tr>
<tr>
<td>Not_error_detected</td>
<td>Syntactic accuracy</td>
<td>64</td>
<td>1.53</td>
<td>0.079</td>
<td>0.114</td>
</tr>
<tr>
<td>Timeliness</td>
<td>Completeness</td>
<td>27</td>
<td>0.65</td>
<td>0.039</td>
<td>0.056</td>
</tr>
<tr>
<td>Timeliness</td>
<td>Syntactic accuracy</td>
<td>17</td>
<td>0.41</td>
<td>0.026</td>
<td>0.038</td>
</tr>
<tr>
<td>Syntactic accuracy</td>
<td>Timeliness</td>
<td>3</td>
<td>0.07</td>
<td>0.006</td>
<td>0.009</td>
</tr>
<tr>
<td>Syntactic accuracy</td>
<td>Not_error_detected</td>
<td>1</td>
<td>0.02</td>
<td>0.002</td>
<td>0.003</td>
</tr>
<tr>
<td>Completeness</td>
<td>Timeliness</td>
<td>1</td>
<td>0.02</td>
<td>0.002</td>
<td>0.003</td>
</tr>
</tbody>
</table>

Table 2. Listing of observed combinations, sorted by descending entropic values

The column Count in Table 2 is an example of the frequency \(N(d_j,d_k,...,d_l)\). In particular, if we consider e.g. for \(X^D\) the timeliness dimension and for \(Y^D\) the completeness dimension, then the number of observations for the combination (timeliness, completeness) is \(N(\text{timeliness, completeness}) = 27\). For two variables \(X^D, Y^D\), the number of combinations is equal to \(L \leq (m+1)^2\), where \(m\) is the number of data quality dimensions and 1 stands for the “Not_error_detected” category. In general, for a set of \(v\) variables \((X^D, Y^D, ..., Z^D)\), the number of combinations is equal to \(L \leq (m+1)^v\). The relative frequencies \(\hat{p}(d_j,d_k,...,d_l)\) are calculated as follows

\[
\hat{p}(d_j,d_k,...,d_l) = \frac{N(d_j,d_k,...,d_l)}{N}
\]

In order to identify, via entropic values, the frequency of error occurrence, for each combination of dimensions, we introduce a variable \(W_{d_j,d_k,...,d_l}\) that takes two values, the first value is equal to the relative frequency \(\hat{p}(d_j,d_k,...,d_l)\) of the combination \((d_j,d_k,...,d_l)\), the second value is equal to \(1 - \hat{p}(d_j,d_k,...,d_l)\). In this way, we obtain \(L\) sub-tables as shown in Figure 2.
The following property detects the combination occurring the most. For two generic combinations, the following relation applies:

$$H_{W(d_j,d_k,...,d_l)}(\hat{p}_{(d_j,d_k,...,d_l)} - 1 - \hat{p}_{(d_j,d_k,...,d_l)}) \geq H_{W(d_j,d_k,...,d_l)}(\hat{p}_{(d_j,d_k,...,d_l)} - 1 - \hat{p}_{(d_j,d_k,...,d_l)}) \Leftrightarrow \hat{p}_{(d_j,d_k,...,d_l)} \geq \hat{p}_{(d_j,d_k,...,d_l)}$$

where the entropic values $H_{W(d_j,d_k,...,d_l)}$, and $H_{W(d_j,d_k,...,d_l)}$ are calculated using definition 1.

The column **Entropic Value**, in Table 2, is an example of $H_{W(d_j,d_k,...,d_l)}$ values. The normalization of the entropic values is expressed by the following formula:

$$E_{W(d_j,d_k,...,d_l)} = \frac{H_{W(d_j,d_k,...,d_l)}}{\text{Log}(2)}$$

where $0 \leq E_{W(d_j,d_k,...,d_l)} \leq 1$.

The list of all the combinations sorted by descending entropic values of the variables $W(d_j,d_k,...,d_l)$ is produced with this framework. In the following section, we address the analysis of dependencies among data quality dimensions.

### 5 Main Models of Dependencies Among Data Quality Dimensions

The focus of this section is the analysis of main models of dependencies for the variables $X^D$ and $Y^D$, and their characterization via entropic equations. The models proposed here to analyze dependencies between two attributes can be extended to a larger number of attributes. Moreover, this analytical framework has been designed for the analysis of data quality dimensions. However parts of the framework can be adapted for the analysis of dependencies among any typology of categorical variables. The dependency models can be classified in three main categories: i) perfect dependency, ii) independency, and iii) partial dependency. For the partial dependency, this framework considers a relevant subset of models: partial synonymy, resolution of $X^D$ is larger than $Y^D$, degeneration of $Y^D$ conditioned by one or more data quality dimensions of $X^D$, and the lower-resolution absolute synonymy.

**Perfect dependency**

The **perfect dependency** model among data quality dimensions of the variable $X^D$ and the variable $Y^D$ occurs when there is a one-to-one correspondence among all the dimensions of $X^D$ and $Y^D$. Figure 3 presents a perfect dependency model, where for example, for the dimension $d_2$ of $X^D$, there is a correspondence to one, and only one, of the dimensions of $Y^D$, in this case $d_1$, and vice versa.
Figure 3. Perfect dependency between $X^D$ and $Y^D$

In the above figures, $N_{jk}$ denotes the cross frequency of the combination $(d_j, d_k)$.

The model of Figure 3 is characterized by the formula (see [22]),

$$H_{X^D|Y^D} (p) = H_{X^D} (p) = H_{Y^D} (p) \neq 0.$$  

**Independency**

There is *independency* between $X^D$ and $Y^D$, when the knowledge of data quality dimensions of $X^D$ does not give any information on the dimensions of $Y^D$. More formally, $X^D$ and $Y^D$ are independent if the probability of joint occurrence is the product of the individual marginal probabilities ($P(X^D = d_j, Y^D = d_k) = P(X^D = d_j) \cdot P(Y^D = d_k)$). The model is identified by the following entropic formulas (see [22]),

$$H_{X^D|Y^D} (p) = H_{X^D} (p) + H_{Y^D} (p)$$

$$I_{X^D|Y^D} (p) = 0.$$

**Partial synonymy between $X^D$ and $Y^D$**

A *partial synonymy* model is defined between $X^D$ and $Y^D$ when the one-to-one correspondence among the dimensions of $X^D$ and $Y^D$ is only partial, i.e., it is defined on proper subsets of $X^D$ and $Y^D$. Figure 4 presents the model of partial synonymy where, for example, for the dimension $d_{m-1}$ of $X^D$, there is a correspondence to one, and only one of the dimensions of $Y^D$, in this case $d_{m-1}$, and vice versa.
To identify the partial synonymy model we propose to use the “collapsing code technique”, used in the statistical domain. For a table of \( m \) rows and \( m \) columns, the collapsing code technique calculates for every cell \((j,k)\), the collapsing of all the observations except for \( X^D = d_j \) and \( Y^D = d_k \). Thus, for each dimension, we define a new variable \( \overline{X}^D_j \), as follows (where \( i \) characterizes the \( i \)-th observation):

\[
\left( \overline{X}^D_j \right) = \begin{cases} 
   d_j, & \text{if } X^D_i = d_j \\
   \overline{d}_j, & \text{if } X^D_i \neq d_j
\end{cases}
\]

Similarly for \( \overline{Y}^D_k \):

\[
\left( \overline{Y}^D_k \right) = \begin{cases} 
   d_k, & \text{if } Y^D_i = d_k \\
   \overline{d}_k, & \text{if } Y^D_i \neq d_k
\end{cases}
\]

Starting from a table of \( m \) rows and \( m \) columns, we compute \( m \times m \) sub-tables of dimension 2×2 as presented in Figure 5 below.

\[
\begin{array}{cccc}
   & Y^D & & \\
   d_1 & d_2 & d_{m-1} & d_m \\
   X^D & N_{11} & N_{12} & 0 & 0 \\
   d_2 & N_{21} & N_{22} & 0 & 0 \\
   ... d_j & N_{j1} & N_{j2} & 0 & 0 \\
   d_{m-1} & 0 & 0 & N_{m-1m-1} & 0 \\
   d_m & 0 & 0 & 0 & N_{mm} \\
\end{array}
\]

**Figure 4. Partial synonymy between \( X^D \) and \( Y^D \)**

**Figure 5. Sub-table related to variables \( \overline{X}^D_j \), \( \overline{Y}^D_k \)**

where \( \hat{p}_{jk} \) is the relative frequency of the combination \( \{d_j,d_k\} \), and \( \hat{p}_{j+} \), \( \hat{p}_{+k} \) are the marginal frequencies. It is possible to detect the model of partial synonymy from the analysis of these sub-tables. If for some \( j,k \) we have that \( H_{\overline{X}^D_j} (\hat{p}) = H_{\overline{Y}^D_k} (\hat{p}) = H_{\overline{X}^D_j, \overline{Y}^D_k} (\hat{p}) \neq 0 \), then the model of partial synonymy has been identified.

**The resolution of \( X^D \) is larger than \( Y^D \)**

The *resolution of \( X^D \) is larger than \( Y^D \)* when there is a one-way perfect dependency (see [22]) of \( Y^D \) from \( X^D \). Figure 6 presents the model of *resolution of \( X^D \) is larger than \( Y^D \)*, where for example, for
the dimension $d_{m-1}$ of $X^D$ there is a correspondence to one, and only one, of the dimensions of $Y^D$, in this case $d_m$. Vice versa is not true, for the dimension $d_m$ of variable $Y^D$ there are two possible corresponding dimensions, in this case $d_{m-1}, d_m$.

$$
\begin{array}{c|c|c|c|c}
& Y^D & \ldots & & \\
\hline
\quad & d_1 & d_2 & \ldots & d_m \\
\hline
X^D & 0 & N_{12} & 0 & 0 \\
& N_{21} & 0 & 0 & 0 \\
& 0 & 0 & N_{jk} & 0 \\
& 0 & 0 & 0 & N_{m-1m} \\
& 0 & 0 & 0 & N_{mm} \\
\end{array}
$$

Figure 6. The resolution of $X^D$ is larger than $Y^D$

This model is identified by the following formulas, (see [22])

$$
H_{X^D Y^D}(p) = H_{X^D}(p)
$$

$$
H_{Y^D}(p) < H_{X^D Y^D}(p) = H_{X^D}(p).
$$

**Degeneration of $Y^D$ conditioned by one ore more data quality dimensions of $X^D$**

This model occurs when for a proper subset of combination $(d_j, d_k)$ there is a one-way perfect dependency between the dimension $d_j$ of $X^D$ and the dimension $d_k$ of $Y^D$. This model is different from the resolution of $X^D$ is larger than $Y^D$ model because the one-way perfect dependency is not true for all dimensions. For example in Figure 7, there is a one-way dependency among dimensions $d_1, d_2, \ldots, d_j, d_{m-1}$ related to $X^D$ and dimension $d_1, d_2, \ldots, d_k, d_{m-1}$ related to $Y^D$.

$$
\begin{array}{c|c|c|c|c|c|c}
& Y^D & \ldots & & & & \\
\hline
\quad & d_1 & d_2 & \ldots & d_k & d_{m-1} & d_m \\
\hline
X^D & 0 & N_{12} & 0 & 0 & 0 & 0 \\
& N_{21} & 0 & 0 & 0 & 0 & 0 \\
& 0 & 0 & N_{jk} & 0 & 0 & 0 \\
& 0 & 0 & 0 & N_{m-1m-1} & 0 & 0 \\
& N_{m1} & N_{m2} & N_{mk} & N_{mm-1} & N_{mm} & \\
\end{array}
$$

Figure 7. The degeneration of $Y^D$ conditioned by one ore more data quality dimensions of $X^D$
Figure 7. Degeneration of $Y^D$ conditioned by one or more dimensions of $X^D$

This model is also characterized by the following formulas:

\[
H_{X^D \rightarrow Y^D} (p) > H_{X^D} (p)
\]

\[
H_{X^D \rightarrow Y^D} (p) > H_{X^D} (p).
\]

**Lower-resolution absolute synonymy**

The lower-resolution absolute synonymy model occurs when there is a perfect dependency among different proper subsets of data quality dimensions. This model is detected using the collapsing code technique, applied to any possible combination of dimensions, as described below. Figure 8 shows an example of the model, where there is a one-to-one correspondence between the following two sets of dimensions $\{X^D = (d_1, d_2, ..., d_j) ; Y^D = (d_1, d_2)\}$ and $\{X^D = (d_{m-1}, d_m) ; Y^D = (d_{m-1}, d_m)\}$.

**Figure 8. Lower-resolution absolute synonymy**

Let $\{d_{i_1}, ..., d_{i_a}\}$ be a generic subset of $\{d_1, d_2, ..., d_m\}$. For the subset $\{d_{i_1}, ..., d_{i_a}\}$ we define the variable

\[
\hat{X}^D_{\{i_1, ..., i_a\}} = \begin{cases} 1 & \text{if } X^D \in \{d_{i_1}, ..., d_{i_a}\} \\ 0 & \text{if } X^D \notin \{d_{i_1}, ..., d_{i_a}\} \end{cases}
\]

Similarly, for the subset $\{d_{k_1}, ..., d_{k_b}\}$ we define the variable

\[
\hat{Y}^D_{\{k_1, ..., k_b\}} = \begin{cases} 1 & \text{if } Y^D \in \{d_{k_1}, ..., d_{k_b}\} \\ 0 & \text{if } Y^D \notin \{d_{k_1}, ..., d_{k_b}\} \end{cases}
\]

From the literature, it is known that a set of cardinality $m$ has exactly $2^m$ subsets (see for example [18]). In this case, we consider the subset of $\{d_1, d_2, ..., d_m\}$ that are not empty and with maximum cardinality equal to $\frac{m}{2}$. The exact number of subsets that we analyze for $\{d_1, d_2, ..., d_m\}$ is

\[
m' = \sum_{j=1}^{\frac{m}{2}} \binom{m}{j} + \frac{1}{2} \left( 1 + m - 2\lambda \right) \binom{m}{\lambda}
\]
where
\[ \lambda = \left\lfloor \frac{m}{2} \right\rfloor \] (i.e. the integer part of \( \frac{m}{2} \)).

For each combination of subsets, a sub-table \( 2 \times 2 \) is defined. The total number of \( 2 \times 2 \) sub-tables is \( m' \times m' \). If among the \( m' \times m' \) sub-tables at least one absolute synonymy exists, then the lower-resolution absolute synonymy between \( X^D \) and \( Y^D \) exists. Formally, we have lower-resolution absolute synonymy when for a pair of subset \( \{f_1, ..., f_n\} \) and \( \{k_1, ..., k_p\} \) the following formula holds
\[
H_{\hat{\tilde{x}}^{\alpha_{\{f_1, ..., f_n\}}}}(p) = H_{\hat{\tilde{y}}^{\alpha_{\{k_1, ..., k_p\}}}}(p) = H_{\hat{\tilde{x}}^{\alpha_{\{f_1, ..., f_n\}} \tilde{y}^{\alpha_{\{k_1, ..., k_p\}}}}}(p) \neq 0.
\]

6 AN APPLICATION OF THE FRAMEWORK ON DATA QUALITY DIMENSIONS RELATED TO FINANCIAL VARIABLES

The analytical framework described in this paper has been implemented in a program, developed in SAS® language, made of approximately 300 lines. The program is available upon request from the authors. This section presents results of the framework application on two variables (Moody’s and Standard & Poor’s bond rating) of a financial database. Before running the program, an a priori data quality assessment has been performed. The technique used in the assessment consists of comparing in-house data with external independent sources in order to detect errors present in the internal data (the external independent source is considered correct). Table 3 below shows the error frequencies and error percentages of detected erroneous observations of the two variables. The total number of analyzed observations is 4175.

<table>
<thead>
<tr>
<th>Data quality dimension</th>
<th>( X^D ) related to attribute rating Moody’s</th>
<th>( Y^D ) related to attribute rating S&amp;P</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Error frequency</td>
<td>Error percent</td>
</tr>
<tr>
<td>Syntactic accuracy</td>
<td>4</td>
<td>0.99</td>
</tr>
<tr>
<td>Completeness</td>
<td>1</td>
<td>0.25</td>
</tr>
<tr>
<td>Timeliness</td>
<td>399</td>
<td>98.76</td>
</tr>
</tbody>
</table>

Table 3. Experimental results of errors distribution: a concrete example

The program produces three different kinds of outputs. Output 1, in Table 4, presents the cross classification of errors on rating Moody’s and rating S&P variables.

<table>
<thead>
<tr>
<th>Data quality dimension</th>
<th>Syntactic accuracy</th>
<th>Completeness</th>
<th>Timeliness</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syntactic accuracy</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>(1.67%)</td>
</tr>
<tr>
<td>Completeness</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>(0.56%)</td>
</tr>
<tr>
<td>Timeliness</td>
<td>17</td>
<td>27</td>
<td>132</td>
<td>(73.33%)</td>
</tr>
<tr>
<td>Total</td>
<td>17</td>
<td>27</td>
<td>136</td>
<td>(75.56%)</td>
</tr>
</tbody>
</table>

Table 4. SAS® program output 1: cross classification table
In order to identify the dependency model that characterizes the frequencies distribution of table 4, the following values are calculated:

\[ H_{X^o}(p) = -0.0167 \cdot \log(0.0167) - 0.0056 \cdot \log(0.0056) - 0.9778 \cdot \log(0.9778) = 0.0517 \]

\[ H_{Y^o}(p) = -0.0944 \cdot \log(0.0944) - 0.15 \cdot \log(0.15) - 0.7556 \cdot \log(0.7556) = 0.3123 \]

\[ H_{X^oY^o}(p) = -0.7333 \cdot \log(0.7333) - 0.15 \cdot \log(0.15) - 0.09444 \cdot \log(0.09444) - 0.01667 \cdot \log(0.01667) - 0.00556 \cdot \log(0.00556) = 0.3613 \]

The data quality dimensions of table 4 are not perfectly dependent. The equation that characterizes the perfect dependency model \( H_{X^oY^o}(p) = H_{X^o}(p) = H_{Y^o}(p) = 0 \) is not true for the above entropic values. Moreover the dimensions are not independent. The equation that characterizes the independency model \( H_{X^oY^o}(p) = H_{X^o}(p) + H_{Y^o}(p) \) does not apply to the above values. The partial dependency model, called in section five degeneration of \( Y^D \) conditioned by one or more data quality dimensions of \( X^D \), applies to the frequencies distribution of table 4. In particular, for the combination \( (d_j = \{\text{syntactic accuracy, completeness}\}, d_k = \{\text{timeliness}\}) \) there is a one-way perfect dependency between the dimension \( d_j \) of \( X^D \) and the dimension \( d_k \) of \( Y^D \). In addition, as reported in section five, the following equations are respected:

\[ H_{X^oY^o}(p) > H_{X^o}(p) \]

\[ H_{X^oY^o}(p) > H_{Y^o}(p) \]

Output 2, in Table 5, presents summary statistics related to the detected model. Output 3, in Table 6, shows the dependency model detected. As listed in Table 6 below, a partial dependency model is defined between \( X^D \) and \( Y^D \).

<table>
<thead>
<tr>
<th>Number of observations</th>
<th>Number of data quality dimensions of variable ( X^D )</th>
<th>Number of data quality dimensions of variable ( Y^D )</th>
<th>Number degeneration of ( X^D ) conditional to ( Y^D = d_k )</th>
<th>Number degeneration of ( Y^D ) conditional to ( X^D = d_j )</th>
</tr>
</thead>
<tbody>
<tr>
<td>180</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 5. SAS® program output 2: summary table of descriptive statistics and models

<table>
<thead>
<tr>
<th>Model</th>
<th>Categories of ( X^D )</th>
<th>Categories of ( Y^D )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degeneration of ( Y^D ) conditional to ( X^D = d_j )</td>
<td>Syntactic accuracy</td>
<td>Timeliness</td>
</tr>
<tr>
<td>Degeneration of ( Y^D ) conditional to ( X^D = d_j )</td>
<td>Completeness</td>
<td>Timeliness</td>
</tr>
</tbody>
</table>

Table 6. SAS® program output 3: detected model of dependencies between \( X^D \) and \( Y^D \)

The application of the analytical framework detected a partial dependency model. Notice, in Table 4, that the most occurring combination of errors among Moody’s and Standard & Poor’s observations is related to timeliness. Errors caused from an inappropriate loading process in terms of timeliness are common to both attributes. We may interpret this main occurrence as due to an inappropriate loading process of the two data sets. This hypothesis has been validated by the real life environment, where the loading process...
of the Moody’s and Standard & Poor attributes in the internal database of the bank, was not adequate to the values frequency updates of the external bond rating data provider. The framework analysis results correctly guided the information analyst to resolve a real problem which occurred on the loading process of bond rating attributes.

7 CONCLUSION AND FUTURE WORK

The analysis of the dependencies among data quality dimensions is extremely important in the area of information quality in order to improve the quality level of a data set, reconstruct the cause-effect patterns on data quality dimensions, select the most important improvement activities, and more generally increase knowledge on dimensions and their relationship. So far, correlation among data quality dimensions has been investigated in the literature focusing the attention on tradeoff analysis, methodology assessment and logical interdependence. In this paper we proposed a data-driven approach for the analysis of dependencies. An analytical framework, defined dependency models terms characterized by entropic equations, has been provided. It is a versatile framework that can be used to discover dependencies among dimensions in the whole assessment, measurement and improvement lifecycle. In order to have an automatic identification of the dependency models, a SAS® program has been implemented. Finally a case study provides evidence of the framework effectiveness. A possible future application could be the enhancing of tradeoff analysis methodologies, through the automatic discovery of functional relationships among dimensions.

In future work, we are interested in extending our research to a wide set of dependencies among dimensions, costs and benefits of data quality in improvement activities. In other words, we are interested in exploring the interaction among these factors: i) the current level of data quality, ii) the planned level of improvement, iii) the existing dependency among dimensions, iv) the actual cost of non quality data v) future savings and benefits due to the achievements of target quality level. In such a way we aim to enhance existing methodologies and approaches for cost benefits analysis in data quality projects (see [4],[7],[16]).

8 ACKNOWLEDGMENTS

We thank Dr. Spyros Arsenis for his contribution to the definition of several analytical concepts that characterise the framework presented in the paper. Many thanks also to Dr. Vasiliki Alexandrou for her careful review of this paper.

9 REFERENCES