Completeness of Attribute Values Representing Partial Information

Fabian Panse
University of Hamburg
Vogt-Kölln Straße 33
22527 Hamburg, Germany
panse@informatik.uni-hamburg.de

ABSTRACT

The data quality dimension *completeness* quantifies the extent to which information on a real-world application is represented in a database. The intensional completeness measure density can be considered at different levels of granularity. With respect to attribute values, current density metrics are undefined for values which represent partial information. Thus, for data models using such concepts these metrics are not suitable and a representative measuring of completeness is not possible. In order to correct this flaw, we redefine the metrics of density at the attribute value level w.r.t. partial information which can be represented by classical subsets of the corresponding attribute domains. Since these redefinitions enable a more accurate and more exact quality measuring of data values, quality based activities and decisions can be more effective in the future.

1. INTRODUCTION

In the relational data model only the null value *null* is available for modeling incomplete information. As a consequence, current metrics of data completeness w.r.t. attribute values consider only two cases: A data value represents complete information on an existing real-world phenomenon (*specified value*) or not (*null*). In real-life applications incomplete information on object properties (for example the age of a person) is not so rare and can appear in different degrees. For a lossless storing of partial information more powerful representation concepts than a single null value are required. Therefore, in the last two decades, several extensions of the relational data model have been proposed: fuzzy models ([11] et al.), probabilistic models ([2] et al.) and models based on other concepts ([3], [6], [9] et al.). In order to measure the increased completeness which can be achieved from using one of these models, current completeness metrics have to be adapted to attribute values representing any degree of information (partial or complete).

In general, we distinguish between information on the existence and information on the occurrence of an object property[1]. While existence can be either known or unknown, information on occurrence has a larger range. The goal of this paper is to adapt completeness metrics to attribute values which represent partial information on existence and/or occurrence. With respect to partial information on occurrences, we focus on such information that can be represented by a classical subset of the corresponding attribute domain (also known as disjunctive- or OR-set ([5], [6])). Treatments for concepts of possibility and probability distributions will be subjects in future work.

In order to demonstrate the necessity of the metrics defined in this paper, we consider a sample database storing personal data. During data storage, the age of a person John Doe is only partially known (between 25 and 30 years). In a relational database, this information can only be stored by the tuple t_1=(John,Doe,*null*). In contrast, by using one of the extended models a lossless storing of the partial information is possible (t_2=(John,Doe,[25,30])). It is obvious that the attribute value $t_2.Age$ represents more information than $t_1.Age$ and hence is more complete. However, $t_2.Age$ contains less information than the *specified value* $t_3.Age$ of the tuple t_3=(John,Doe,28). Thus, the completeness of $t_2.Age$ has to be between the completeness of $t_1.Age$ and the completeness of $t_3.Age$. As a consequence, special metrics for measuring completeness of values representing partial information are required.

We think, besides completeness measuring in the context of quality assessment or quality improvement activities, the redefined metrics can be used in different application areas. For instance, in order to quantify the information loss resulting from a data anonymization. This in turn helps to balance between the two contrary goals de-identification and practical usefulness of the anonymized data.

The paper is structured as follows: Section 2 examines related work and presents current metrics of data completeness especially of data density. In Section 3 the different degrees of information which can be available on an object property are analyzed and classified into six representing information classes. In Section 4 we define new density metrics for attribute values representing information of the previously defined classes w.r.t. four different kinds of attribute domains: countable and finite (e.g. the color of a car), countable and indefinite (e.g. the age of a car), uncountable and bounded (e.g. the fulness of the petrol tank in percentage) and last but not least uncountable and unbounded (e.g. the mileage of a car). In order to demonstrate

[1]The occurrence is the value of an existing object property (e.g. the occurrence of 'QDB 2009's venue' is 'Lyon').

the newly defined metrics we present an example of calculating densities in Section 5. Usually attribute domains do not exactly represent the scope of the corresponding object property (e.g. no car is driven a million miles). Thus, besides simple metadata (e.g. the attributes' domains themselves) additional information on the corresponding application domain can be used to increase the significance of the resulting completeness values. Two kinds of such information, namely *domain restricting knowledge* and *probability distributions on attribute domains*, are shortly discussed in Section 6. A final conclusion summarizes the paper and gives an outlook on future work.

2. RELATED WORK

Metrics of data completeness are considered in different works (Scannapieco ([10]), Naumann ([8]), Motro ([7]) et al.), but none of them regard the existence of partial information at the level of attribute values.

Naumann defines completeness from an extensional (data coverage) and an intensional (data density) point of view. While the coverage of a relation is the ratio of all stored to all actually existing entities of the modeled entity type, data density is the completeness of the stored entities and can be considered at different granularities (e.g. relation, tuple and attribute value).

Since only the intensional completeness is affected by the capability of representing partial information in attribute values, in the following we exclusively relate to the intensional completeness measure density. In the relational data model an attribute value is either a single element of the corresponding domain (*specified value*) or the null value *null* which represents the case that no information is available. Consequently, the density of an attribute value is either 1 (*specified value*) or 0 (*null*). The density of a tuple or a relation is the average density of its attribute values or tuples respectively. In order to adjust completeness on individual application domains, for tuple density each attribute A can be rated by a weight-value $w_A \in [0,1]$. Given a single relation $\mathcal{R} = (A_1, ..., A_n)$ the density of a tuple $t \in \mathcal{R}$ and the density of $\mathcal{R}$ are defined as follows:

$$d(t) = \frac{\sum_{i=[1,n]} w_{A_i} d(t.A_i)}{\sum_{i=[1,n]} w_{A_i}} \quad (1) \qquad d(\mathcal{R}) = \frac{\sum_{t \in \mathcal{R}} d(t)}{|\mathcal{R}|} \quad (2)$$

With respect to partial information the density of an attribute value has to be defined as a value in the range $[0,1]$ instead only 0 or 1. The densities of tuples and relations can be calculated as usual.

3. INFORMATION CLASSIFICATION

In general, information on an object property can be divided into two categories which can be considered independently to a large extent: Information on the existence (I_E) and information on the occurrence (I_O) of this property. The existence of a property is either known or not. Thus information of this kind can be separated into two classes. In contrast, information on occurrence cannot be categorized in discrete classes, but is within a continuous range between the total ignorance (no information on the occurrence is available) and the total knowledge (the occurrence is exactly known). The whole *spectrum of information* is shown in Figure 1.

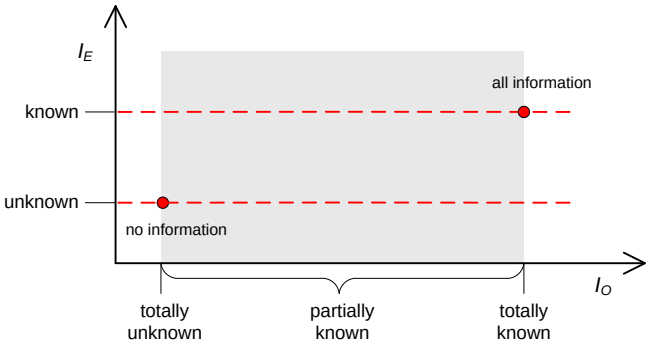

Figure 1: spectrum of information

If we divide the continuous range of occurrence information into three areas, two for the extremes totally unknown and totally known and one for the range between them (partially known), six combinations of information on existence and information on occurrence (in the following denoted as information classes) are possible:

1. no information: Since neither information on existence nor information on occurrence is available, this class represents total ignorance.

2. no existence but partial occurrence information: Sometimes, it is not known if an object property exists or not, but if the property is existent, some partial information on its occurrence is available. For example, it is unknown if John has a phone, but his adress (Hamburg) is well known. Thus, if he has a phone at least the dialing code (040) of the phone number is known.

3. no existence but complete occurrence information: As for the elements of class 2, it is not known whether the corresponding property is applicable for the considered object or not. However if it is applicable, than its occurrence information is well known (see the example of the driver license in Section 5).

4. existence but no occurrence information: In many cases, it is known, that an object property exists, but its occurrence is totally unknown (e.g. the unknown age of a person).

5. existence and partial occurrence information: The members of this class contain the information that the object property exists, but on its occurrence only partial information is available (e.g. the age of a person is known to be between 20 and 30 years).

6. all information: In this case, all information (existence as well as occurrence) is known. This applies, if either it is known that the property exists and exact information on its occurrence is available (e.g. the age of a person is exactly known as 25 years) or it is known that the property does not exist (e.g. a person has no driver license). In the latter case, it is obvious that no occurrence exists, too.

Since information on existence and information on occurrence cannot be compared without any measure, this set of classes is only partial ordered (see the Hasse diagram in Figure 2).

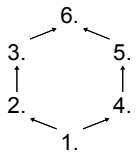

Figure 2: Hasse diagram of information classes 1-6

In order to represent the different kinds of partial information in some works null values ([1], [3] et al.) are used. In general, if we neglect reasons for information absences all kinds of null values which are defined in previous works can be assigned to one of these information classes (e.g. since from the null value "not allowed to be read" no information can be derived, this null value has to be assigned to information class 1).

4. DENSITY OF ATTRIBUTE VALUES

As pointed out in the previous section, we distinguish information on existence and information on occurrence. As a logical consequence an attribute value's density can also be decomposed into two parts, one density for the existence information (d_E) and one density for the occurrence information (d_O).

Since the relative importance of existence information compared to occurrence information depends on the corresponding application domain, the two weight values $w_E \in [0,1]$ and $w_O = 1 - w_E$ are introduced and can be variably defined for the individual attributes. For example, since every person has an age, for the attribute Age the existence information is implicated by the attribute's semantics and the existence is weighted by $w_E = 0$. In some domains, nonexistence can be modeled by a single domain element (e.g. no salary is represented by the value 0). In this cases, a separate consideration of existence information is needless and can be avoided by setting $w_E = 0$. Furthermore, for the middle name of a person information on existence is relative worthless without any information on occurrence. Consequently, for the corresponding attribute a weight-value $w_O >> w_E$ has to be chosen. In contrast, w.r.t. the driver license, often it is more important to know, if an employee is allowed to drive a car than the information on what kind of car he is allowed to drive ($w_E >> w_O$).

In general, the introduction of these two weight-values enables an application domain depending ordering of the six information classes proposed in the last section. Hereafter, the density of an attribute value v is defined as:

$$d(v) = w_E \cdot d_E(v) + w_O \cdot d_O(v) \tag{3}$$

4.1 Density of Existence Information

The density d_E quantifies whether the value contains information on the existence of the corresponding object property or not. Since the existence of a property can either be known or unknown, the density d_E can only be 1 (if existence is known) or 0 (if not). The first case concerns the information classes 4-6 and the second one the classes 1-3.

$$d_E(v) = \begin{cases} 1 & \text{existence information is available} \\ 0 & \text{else} \end{cases}$$

4.2 Density of Occurrence Information

The density d_O quantifies to what extent the value contains information on the occurrence of the corresponding object property. As already mentioned, in contrast to the bivalent existence information, the occurrence information is within the continuous range between the two extremes totally unknown and totally known. From attribute values which represent information of the classes 3 and 6 complete knowledge on the property's occurrence can be derived. Thus, these values are at the one end of the scale (totally known) and have the density $d_O = 1$. In contrast, attribute values which represent information from the classes 1 and 4 are at the other end of the scale (totally unknown) and have a density $d_O = 0$. All the other values (values which represent partial occurrence information) are between these two extremes and have a density $d_O \in (0,1)$.

Usually, the partial information represented by an attribute value v can be used to reduce the corresponding attribute domain D_A to a subset containing all elements which are still a possible occurrence of the corresponding property. In the following this subset is denoted as the partial set $S_V \subseteq D_A$. In the extreme cases, the partial set is either equal to the domain ($S_V = D_A$, totally unknown) or consists of just one element ($|S_V| = 1$, totally known)[2]. All elements of the partial set are considered to be equally plausible to be the true occurrence. Thus, for simplification purposes, information on correlations between single elements (e.g. different probabilities) are neglected at the moment (see future goals in Section 7).

In order to satisfy the metric requirements of interval scale and interpretability ([4]), the simplest metric $d_O(v) = 1/|S_V|$ is not really a choice. A more adequate metric can be achieved, if the size of the partial set is related to the corresponding attribute domain's size. The more elements of the domain were already be excluded, the smaller is the partial set and the higher is the resulting density. In general, the domains of the individual attributes can be divided in four categories, each described in one of the following subsections.

4.2.1 Countable and Finite Domains

A set S is countable, if there exists a bijective mapping of S to the set of the natural numbers $\mathbb{N}$. In contrast to $\mathbb{N}$ the domains of this category are finite. For countable sets, the cardinality $|S|$ of a set S is defined as its total number of elements. Using the cardinality, an appropriate metric for the density d_O of an attribute value v can be achieved from the ratio of the number of elements which already have been excluded ($|D_A| - |S_V|$) to the total number of domain elements which have to be totally excluded ($|D_A| - 1$) in order to get a single element. Since the density of an attribute value that represents a nonexistent object property (no occurrence exists and S_V is an empty set) has to be the value 1, additionally the minimum operator has to be applied.

$$d_O(v) = min(1, \frac{|D_A| - |S_V|}{|D_A| - 1}) \tag{4}$$

Since a feasible domain contains at least two elements (denominator is not 0), an invalid result can be excluded. Due to $S_V \subseteq D_A$, the resulting ratio is always in the range $[0,1]$ and hence d_O is normalized. Furthermore, the metric of

[2]or 0 elements if the property does not exist.

equation 4 is interval scaled and interpretable: The density d_O is 1/2, if the number of already excluded elements is equal to the number of elements which have to be further excluded in order to get a *specified value*. Besides normalization, interpretability and interval scale all other metric requirements listed in [4] are also satisfied.

4.2.2 Countable and Infinite Domains

The power of a countable and infinite set S (e.g. $S = \mathbb{Q}$) is equal to the power of $\mathbb{N}$. Since the metric of equation 4 converges to 0 if $|D_A|$ and $|S_V|$ converges to infinity, for a countable and infinite domain this metric is not useable.

Interpretability and interval scale require a consideration of the cardinality of both, the partial set as well as the attribute's domain. In this context, a specification of an adequate metric w.r.t. a domain which contains an infinite number of elements is impossible.

However, in most cases, the scope of the corresponding application domain is actually not infinite. Thus, the number of domain elements can be mostly restricted to a finite subset without any quality loss. In order to get adequate boundaries (e.g. no person is older than 130 years or taller than 2.40 m) those restrictions have to be defined by domain experts. If such an expert is not available, database statistics can be used. The minimal/maximal value of the associated attribute stored in the database can act as a lower/upper bound. The better the stored attribute values represent the actual scope of the modeled application domain the more effective such an approach is. Therefore, domain restrictions by statistical values are most practicable, if the corresponding attributes contain a sufficient number of values which are wide spread on the associated application domains (for example in the range 6-20 for the age of schoolchildren).

In the following, a restriction on the interval $[l, u]$ (lower bound l and upper bound u) reduces a countable and infinite set S to the countable and finite set:

$$S^{[l,u]} = \{e | e \in S \wedge e \in [l, u]\} \tag{5}$$

For avoiding a needless information loss, existing bounds of the two sets D_A and S_V have to be incorporated into the composition of l and u. For that purpose, the two operations $min_L()$ and $max_L()$ are required:

$$min_L(S) = \begin{cases} \infty, & min(S) \to -\infty \\ min(S), & \text{else} \end{cases} \tag{6}$$

$$max_L(S) = \begin{cases} -\infty, & max(S) \to \infty \\ max(S), & \text{else} \end{cases} \tag{7}$$

Given $min(R(A))$ and $max(R(A))$ as the minimal- and maximal value stored in the attribute A of the relation $\mathcal{R}$, the two bounds l and u are calculated as:

$$l = min(min(\mathcal{R}(A)), min_L(S_V), min_L(D_A))$$
$$u = max(max(\mathcal{R}(A)), max_L(S_V), max_L(D_A))$$

Using these restricitions, the density d_O results in:

$$d_O(v) = min(1, \frac{|D_A^{[l,u]}| - |S_V^{[l,u]}|}{|D_A^{[l,u]}| - 1}) \tag{8}$$

In general, the larger the restricted scope, the lower is the fault caused by the boundary. Since distortions cannot be completely avoided, the existence of infinite domains should be limited at a minimum by adequate restrictions defined during database design (e.g. CHECK age < 130).

4.2.3 Uncountable and Bounded Domains

With respect to domains of databases, uncountable sets are mostly ordered (the set of the complex numbers is uncountable and non-ordered, but is also not a typical database domain). An uncountable and bounded domain is an uncountable set with a lower and an upper bound (e.g. $\{e | e \in \mathbb{R} \wedge e \in [0, 10]\}$).

Since a continuous and uncountable set $S = [S_{min}, S_{max}]$ contains an infinite number of elements, its cardinality is defined as the distance between the minimal and maximal element of the interval ($|S| = S_{max} - S_{min}$). The cardinality of an uncontinuous and uncountable set (e.g. $\{e | e \in \{[0, 2] \cup [4, 5]\}\}$) results from the sum of the cardinalities of its disjunct and continuous subsets.

$$|S| = \sum_{i=1}^{n} |S_i| \text{ where } (\forall i, j \in [1, n]) : S_i \cap S_j = \emptyset \wedge \bigcup_{i=1}^{n} S_i = S$$

In the simplest case, the partial set as well as the attribute's domain contain only uncountable subsets and the density d_O can be calculated as[3]:

$$d_O(v) = \frac{|D_A| - |S_V|}{|D_A|} \tag{9}$$

Since the cardinality of an interval $[a, a]$ is 0, for a *specified value* the density $d_O = 1$ results. Does the set S_V contain countable subsets, the metric of equation 9 becomes unrepresentative and specific approximations have to be used. Nevertheless, this aspect is out of the scope of this paper and will be considered by further research.

4.2.4 Uncountable and Unbounded Domains

If an uncountable set is unbounded, its cardinality converges to infinity. Thus, as for countable and infinite domains, a restriction is required. By using the same concept of boundary composition as for countable domains, the density d_O results in:

$$d_O(v) = \frac{|D_A^{[l,u]}| - |S_V^{[l,u]}|}{|D_A^{[l,u]}|} \tag{10}$$

5. DEMONSTRATING EXAMPLE

In order to demonstrate the density metrics defined in the last section we consider a part of the relation *employee* of a company's database (see Table 1 and 2). In the relation, the first name (attribute FName), the surname (SName), the age (Age), the class of driver license (DLicense) and the salary per hour (Salary) are stored. Furthermore, for reasons of safety, each employee has to take part on an annual first aid test. The results (in percentage) from the latest test are stored in the attribute FirstAid.

The attribute Age is defined in the integer domain and hence is countable and infinite. The domain of the attribute DLicense is a self defined domain (countable and finite) consisting of three elements: the possible driver license classes A, B and C. The attributes FirstAid and Salary are

[3]Since an empty set has the cardinality 0, for information on nonexistence the density $d_O = 1$ correctly results without using the minimum operator.

	FName:String	SName:String	Age:Int	DLicense:D_4	FirstAid:D_5	Salary:Real
	$w_E = 0$	$w_E = 0$	$w_E = 0$	$w_E = 0.7$	$w_E = 0.2$	$w_E = 0$
	...	...	...	...	...	...
t_1	Ralph	Marshall	17	nE	76.67	pk_1
t_2	Tyler	Corman	pk_2	A	npk_1	21.78
t_3	Lisa	Torres	31	npk_2	57.13	27.41
t_4	Dave	Conroy	uk	nek_1	$null$	41.53
	...	...	...	...	...	...

Table 1: a part of the relation *emplyoee*

$D_5 =$ Real CHECK FirstAid BETWEEN 0 AND 100

$D_4 = \{A, B, C\}$ $\quad pk_1 = \{[8, 12] \cup [15, 17]\}$
$pk_2 = [20, 30]$ $\quad npk_1 = [0, 37.5]$
$npk_2 = \{A, B\}$ $\quad nek_1 = \{A\}$

Table 2: domains and null values

defined in the domain of the real numbers (uncountable). Whereas the domain of the attribute Salary is unbounded, the domain of the attribute FirstAid is restricted to the interval $[0, 100]$.

Some newly hired employees have not yet passed a first aid test. Thus, for some employees no test results exist. In general, information on the test result are quite more important than information on its existence ($w_E = 0.2$). Since a nonexistent salary can be represented by the value 0, the weighting w_E is 0 for this attribute.

In order to store the different degrees of information (existence as well as occurrence) in the database the null value concept is used. We take the same five null value semantics as known from [1] and [3], and add a sixth null for representing the case, that existence information is unknown and occurrence information is completely known (class 3). Thus, beside the *specified value*, the null values *no information* (*null*), *maybe existent and partially known* (*npk*), *maybe existent and known* (*nek*), *existent but unknown* (*uk*), *existent and partially known* (*pk*) and *not existent* (*nE*) are defined for the information classes 1-6 in exact this order.

Since some information is incomplete, a couple of attribute values correspond to some of these null values. A special case of incomplete information relates to the class of driver license of tuple t_4. Actually, this person has a driver license of class A, but in the last months, he had a car accident and the license has been confiscated by the police. The responsible secretary does not know, if the confiscation was just for a month or longer. Thus, at the last time of data update, it was not clear, whether this employee has a driver license or not, but if he has one, then this license is definitely of class A.

The densities of the attribute values which contain one of the two null values *null* and *uk* is always 0 or w_E respectively. The densities of the other attribute values which do not contain a *specified value* is calculated as follows:

- Since for this attribute value all information is known ($nE \Rightarrow S_V = \emptyset$), the density of t_1.DLicense results in:

$$d(t_1.\text{DLicense}) = 0.7 \cdot 1 + 0.3 \cdot min(1, \frac{3-0}{2}) = 1$$

- The salary of the employee Marshall (t_1) is only partially known ($pk_1 = \{[8, 12] \cup [15, 17]\}$). By definition, the domain of the salary is unbounded, but domain experts know, that there exists a minimum salary $min(Salary) = 5.00$, which is required by law, as well as an in-house salary cap $max(Salary) = 50.00$. Using these boundaries, the density of t_1.Salary results in:

$$\begin{aligned} d(t_1.\text{Salary}) &= 0 + 1 \cdot \frac{|\mathbb{R}^{[5,50]}| - |\{[8, 12] \cup [15, 17]\}|}{|\mathbb{R}^{[5,50]}|} \\ &= \frac{45-6}{45} = 0.87 \end{aligned}$$

- The employee Cormann (t_2) has recently started a practical work in the company. The responsible secretary does not know if he was already a trainee when the last first aid test took place. Thus, it is uncertain if for this person a test result exists, but it is known that in the last test, no newbie had more than 37.5 percentage ($npk_1 = [0, 37.5]$). Since the corresponding domain is an uncountable and bounded set, the metric defined in 9 has to be used for calculating the density:

$$d(t_2.\text{FirstAid}) = 0.2 \cdot 1 + 0.8 \cdot \frac{100 - 37.5}{100} = 0.625$$

- The age of employee Cormann (t_2) is also only partially known ($pk_2 = [20, 30]$). The corresponding domain is a countable and infinite set. Thus, before calculating the density the domain has to be restricted to a bounded subset. If we assume that no domain expert is available, but the database statistic supplies the minimal (16) and the maximal (64) value of this attribute, the density is calculated as:

$$\begin{aligned} d(t_2.\text{Age}) &= 0 + 1 \cdot min(1, \frac{|\mathbb{Z}^{[16,64]}| - |[20, 30]|}{|\mathbb{Z}^{[16,64]}| - 1}) \\ &= min(1, \frac{49-11}{48}) = 0.79 \end{aligned}$$

- It is not known whether employee Torres (t_3) has a driver license or not, but it is known that she is definitely not allowed to drive a truck (license class C) ($\Rightarrow npk_2 = \{A, B\}$). Therefore, the density of the attribute value t_3.DLicense results in:

$$d(t_3.\text{DLicense}) = 0.7 \cdot 0 + 0.3 \cdot min(1, \frac{3-2}{3-1}) = 0.15$$

- As mentioned above, it is unknown if employee Conroy (t_4) currently has a driver license, but if he has one, its class is exactly known ($nek_1 = \{A\}$). Thus, the density of the corresponding attribute value results in:

$$d(t_4.\text{DLicense}) = 0.7 \cdot 0 + 0.3 \cdot min(1, \frac{3-1}{3-1}) = 0.3$$

If we assume, that all attributes have an equal importance ($(\forall A \in \{\text{FName}, \text{SName}, ..., \text{Salary}\}) : w_A = 1/6$), the densities of the individual tupels result in:

$$d(t_1) = 0.98 \quad d(t_2) = 0.90 \quad d(t_3) = 0.86 \quad d(t_4) = 0.55$$

Consequently, the density of the subrelation *employee'*$= \{t_1, t_2, t_3, t_4\}$ is:

$$d(employee') = \frac{0.98 + 0.90 + 0.86 + 0.55}{4} = 0.82$$

6. FURTHER REMARKS

Besides information that can be derived from metadata (e.g. the number of domain elements) additional information, for example resulting from the knowledge of domain experts, can be used to enhance the significance of the calculated quality values. Although further kinds of information may exist, in the following we only focus on so called *domain restricting knowledge* and *probability distributions*.

6.1 Domain Restricting Knowledge

In many cases, attribute domains do not exactly represent the possible occurrences of the modeled entity types. For example, the home country of a person is defined as a string, but not every string does refer to an existing country. Thus, additional information can be used to restrict the corresponding domain to a large extent. At the moment, the UNO listed 193 sovereign states. Thus, if we assume that a string is coded in ASCII (94 printable characters) and the length of a string is restricted on 50 characters, the number of possible occurrences for the values of an attribute HomeCountry which is defined in the domain String can be significantly decreased from $94^{50} = 4.53 \cdot 10^{98}$ to 193.

6.2 Probability Distributions

Usually, in many applications the frequencies of the individual domain elements are not equal and some elements occur more often than other ones (e.g. there exist more 40 years old persons than persons which are 120 years old). To represent this imbalance, probability distributions can be defined on the corresponding domains. If information on such a distribution is available, instead of the partial set's cardinality, the probabilities of its elements can be used for calculating the density. For example, given a probability density function $f_A(x)$ on the uncountable domain D_A ($P(D_A) = 1$), for a value v representing the interval $[a, b]$, the density d_O can be defined as:

$$d_O(v) \quad = \quad \frac{P(D_A) - P(S_V)}{P(D_A)} = 1 - \int_a^b f_A(x)dx \quad (11)$$

Since in uncountable domains the probability of a single domain element e is $P(e) = 0$, for a known occurrence still the density $d_O = 1$ results. In probability distributions on countable sets, a single element can have a probability greater than 0. Thus, w.r.t. countable domains the probability of a partial set can be lower than the probability of a single element. Since a value representing a single element has the maximal density, the fundamental principle of this approach meaning that the probability and the density are always inversely proportional to each other can be violated. This is a problem which cannot be solved easily and requires further considerations.

7. CONCLUSION

Existing metrics for data completeness at the attribute value level are only defined for two kinds of values: *specified values* which represent complete information and the null value *null* which represents no information on the modeled object property. Since there is a wide range of partial information whose completeness lies between these two extremes, metrics for attribute value completeness have to be adapted to concepts representing such cases.

In order to realize such an adaption, we have adopted Naumann's concept of decomposing completeness into coverage and density and have extended the density at the attribute value level w.r.t. different kinds of partial information. To satisfy several requirements for an adequate quality metric, the partial information has to be related to the corresponding attribute domain. Therefore, we have considered the partial information as a domain subset and have defined the density of an attribute value as the ratio of the number of domain elements which are already excluded to the number of domain elements which have to be totally excluded to get a *specified value*. Usually, attribute domains are either countable and finite, countable and infinite, uncountable and bounded or uncountable and unbounded. In this paper, all four cases have been regarded and we have proposed a possible solution for each of them.

For simplification, we only have considered partial information which can be represented by classical subsets (disjunctive sets) of the corresponding domains. Nevertheless, there exist further concepts for modeling incomplete information like possibility (fuzzy databases) or probability distributions (probabilistic databases). Thus, in order to enable the measurement of completeness in such databases, the basic approach proposed in this paper has to be extended. Furthermore, we have mentioned the problems which arise when relating countable partial sets to uncountable domains. This is also a problem that has to be tackled in future work.

8. REFERENCES

[1] P. Atzeni and V. D. Antonellis. *Relational Database Theory*. Benjamin/Cummings, 1993.

[2] D. Barbará et al. The Management of Probabilistic Data. *IEEE Trans. Knowl. Data Eng.*, 4(5):487–502, 1992.

[3] K. Candan et al. A Unified Treatment of Null Values Using Constraints. *Inf. Sci.*, 98(1-4):99–156, 1997.

[4] B. Heinrich et al. Metrics for Measuring Data Quality - Foundations for an Economic Data Quality Management. In *ICSOFT (ISDM/EHST/DC)*, pages 87–94, 2007.

[5] T. Imielinski. Incomplete Information in Logical Databases. *IEEE Data Eng. Bull.*, 12(2):29–40, 1989.

[6] Z. Michalewicz and L. Groves. Sets and Uncertainty in Relational Databases. In *IPMU*, pages 127–137, 1988.

[7] A. Motro and I. Rakov. Estimating the Quality of Databases. In *FQAS*, pages 298–307, 1998.

[8] F. Naumann et al. Completeness of Integrated Information Sources. *Inf. Syst.*, 29(7):583–615, 2004.

[9] A. Ola et al. Incomplete relational database models based on intervals. *IEEE Transactions on Knowledge and Data Engineering*, pages 293–308, 1993.

[10] M. Scannapieco and C. Batini. Completeness in the relational model: a comprehensive framework. In *IQ*, pages 333–345, 2004.

[11] M. Umano et al. Fuzzy Relational Algebra for Possibility-Distribution-Fuzzy-Relational Model of Fuzzy Data. *J. Intell. Inf. Syst.*, 3(1):7–27, 1994.