APPLYING PROBABILISTIC MODELS TO DATA QUALITY CHANGE MANAGEMENT

Adriana Marotta and Raúl Ruggia
Universidad de la República, Montevideo, Uruguay

Keywords: Data Quality, Data Integration Systems, Changes.

Abstract: This work focuses on the problem of managing quality changes in Data Integration Systems, in particular those which are generated due to quality changes at the sources. Our approach is to model the behaviour of sources and system data quality, and use these models as a basic input for DIS quality maintenance. In this paper we present techniques for the construction of quality behaviour models, as well as an overview of the general mechanism for DIS quality maintenance.

1 INTRODUCTION

Data Integration Systems (DIS) integrate data from a set of heterogeneous and autonomous information sources and provide it to users. Users access this data through queries over the sources as well as over an integrated schema. Due to the increasing use of this kind of systems and their increasing amount of information and sources, the problem of data quality has gained a great importance (Scannapieco, 2005). Furthermore, considering the autonomy and heterogeneity of the data sources, it cannot be ignored that the quality of the DIS may be constantly changing, strongly affecting the information received by the user and the decisions that are made from it. For example, an enterprise manager that queries a Data Warehouse (DW) about the sales of his company should always be aware of the quality of the information he uses. On the other hand, the system should be capable of satisfying his information quality needs.

The DIS basically consists of a set of *data sources*, a *transformation process* that is applied to data extracted from the sources, and a set of *data targets*, which may be the result of pre-defined queries or part of an integrated schema. Over this system, *quality factors* are considered, distinguishing between quality-factors values provided by the system and quality-factors values required by the users. Values are associated to the sources and to the data targets. We focus on *freshness* and *accuracy* factors.

This work addresses the problem of managing DIS quality changes, in particular the changes that are generated due to quality changes at the sources.

Quality values at the sources may be continuously changing and also DIS quality. Our approach is to model the behaviour of sources quality and DIS quality, and to use these models as a basic input for DIS quality maintenance.

The proposed quality behaviour models are probabilistic models of the quality values, and they represent the probabilities of the occurrence of such values. For example, the model of a source's accuracy will provide the probability of each possible accuracy value of the source. This is a prediction of the quality of the source in the short term. In the same way, the quality of the overall DIS can be modelled based on the sources quality models, also considering other system properties.

The proposed quality models contribute to implement tolerance to source quality changes in DIS. If the DIS is designed considering the probabilities of the quality values so that it satisfies user quality requirements, there will be many quality changes that will be endured by the system because they are under the predictions of the models. Therefore, the DIS will not be affected at all by these quality changes.

The main advantages of this approach are the following: (i) predictions about quality behaviour allow the DIS to tolerate source quality changes, (ii) modifications on the DIS caused by quality changes are minimized, and (iii) knowledge about quality

Marotta A. and Ruggia R. (2008).
APPLYING PROBABILISTIC MODELS TO DATA QUALITY CHANGE MANAGEMENT.
In *Proceedings of the Third International Conference on Software and Data Technologies - ISDM/ABF*, pages 296-299
DOI: 10.5220/0001890302960299

behaviour is used for the determination of improvement actions to apply to the DIS.

Our goal in this paper is to present the proposed techniques for the construction of quality behaviour models.

Section 2 presents sources quality behaviour models, Section 3 presents DIS quality behaviour models and Section 4 presents the conclusions.

2 QUALITY BEHAVIOR MODELS OF SOURCES

We model quality behaviour through probabilistic models where we define:

Random Experiment: verification of a source quality value.

Random Variable (RV): source quality factor value.

Sample Space: set of all source quality possible values.

The source model provides us the probability distribution of the quality values at the source, and useful indicators such as expectation, mode, maximum, minimum. This model allows knowing the probability that holds for each of the possible quality values of the source if we query it at any moment.

We build the models through three possible mechanisms, depending on the characteristics of the quality factor and the system, and on the available information. The first one is using an existing distribution. This can be done if the behaviour of the quality factor or some property, on which it is based, is already represented in a theoretical model. The second one is calculating the distribution. This can be done if we have enough information about the behaviour of the quality factor for deducing how the probabilities distribute across the possible values. The third one is obtaining the probability distribution through the utilization of statistical techniques. We build relative frequencies tables (Canavos, 1984) with the collected values of the quality factor. These tables give the relative frequency of each value, which are a good estimation of the respective probabilities (Canavos, 1984) (Cho, 2003).

2.1 Freshness Quality Models

In the following we show a probabilistic model we constructed for a particular scenario.

Model. In this model the sources loading is in a continuous basis. We assume that source updates follow a Poisson distribution and update frequency λ is available. We build the model deducing the probability distribution of freshness values from the distribution of the updates in the source.

Given a source S, the RV X represents the quantity of updates in a time unit, the RV Y represents the source freshness. We know the distribution of X, we deduce the distribution of Y.

The probability that there is at least one update in a certain time interval (p_U) is the complement of the probability that there is zero updates in it:

$$p_U = 1 - p(X=0) \quad (1)$$

The probability that the freshness at the end of certain time interval is 0, is equal to p_U. The probability that the freshness is 1, is equal to the probability that there has been an update in the previous time interval, multiplied by p_U. In this way we can obtain the distribution for the RV Y:

$$\begin{aligned} &p(Y=0) = p_U \\ &p(Y=1) = p_U.p(X=0) \\ &p(Y=2) = p_U.p(X=0).p(X=0) \\ &\ldots\ldots \end{aligned} \quad (2)$$

We calculate the expectation:

$$E(Y) = \Sigma_y\, yp(y) = p(X=0)+(p(X=0))^2+... \quad (3)$$

3 QUALITY BEHAVIOR MODELS OF DIS

We model DIS quality behaviour through three different perspectives: (i) probability of satisfying a quality requirement (which we call *DIS Quality Certainty*), (ii) probability distribution of possible quality values provided by the DIS, and (iii) satisfaction of quality requirements about averages, most probable values, and maximums /minimums. While (i) and (iii) characterize DIS quality considering the existing user quality requirements, (ii) is aimed to show the quality provided by the DIS regardless of the quality requirements.

For (i) and (iii) we need to previously calculate the required quality values at the sources, called *Accepted Configurations*, which are deduced from the user quality requirements.

In the following sub-sections we present the *Accepted Configurations*, the definition and calculation of *DIS Quality Certainty*, and the calculation of DIS quality probability distribution and of the satisfaction of user quality requirements.

3.1 Accepted Configurations

From user quality requirements we deduce the required values at the sources, i.e. the quality restrictions the sources must satisfy. As there may be different combinations of sources quality values for a quality factor that satisfy these restrictions, we call them *accepted configurations*, and we specify them in the following way:

- Source restriction r. A restriction which must be verified by the values of the quality property qp of the source relation R.

$$r = qp(R) \text{ op } n, \quad \text{where op may be } <, \leq, >, \geq, \text{ or } = \qquad (4)$$

- Restriction vector v. A set of restrictions, each one of a source relation.

$$v = <r_1, \ldots, r_n> \qquad (5)$$

- Restriction-vector Space vs. A set of restriction vectors.

$$vs = \{v_1, \ldots, v_m\} \qquad (6)$$

By means of the propagation of the user quality required values to the sources we calculate the *restriction-vector space*, which contains the accepted configurations.

3.2 DIS Quality Certainty

According to the Reliability theory (Gertsbakh, 1989), “the word reliability refers to the ability of a system to perform its stated purpose adequately ... under the operational conditions encountered”. In particular, *structural reliability* relates the state of the system to the state of its components, and gives the probability that the system is operational. The components and the system have two possible states: “operational” and “failure”.

We make the analogy with this definition in the following way: our system is the DIS, our components are the data sources, and the system is operational when its quality requirements are being satisfied. We define *DIS Quality Certainty* as the probability of DIS quality requirements satisfaction.

In our case the state of the system is also dependant on the state of its components. But when are our components operational? We study the case of freshness and the case of accuracy.

In the case of **freshness** each source is operational when it satisfies the corresponding quality restriction from the accepted configurations. We consider the system as a “series system”, which implies that the system is operational if and only if all of its components are operational. We start calculating for each source, the probability of being operational. For each source we consider the random experiment defined for source quality models. The variables determining the component state are binary RVs. Considering n sources, for the restriction vector $v=<r_1,\ldots,r_n>$ we have a set of n RVs Y_i, i=1..n, each of which is equal to 1 if the quality value at source i satisfies the restriction r_i, and otherwise is equal to 0. We obtain the distribution of Y_i from the probability models of the sources as follows. Given that r_i = (freshness$_i$ ≤ k_i), the probability that source i verifies restriction r_i is $p_i = p(Y_i = 1) = p(X_i \leq k_i)$, where X_i is the RV of source freshness. Then, $p_i = p(X_i = 0) + p(X_i = 1) + \ldots + p(X_i = k_i)$.

The probability of being operational for series systems (reliability) is calculated as $r = \Pi_{i=1..n}\, p_i$, where p_i is the probability that component i is operational.

We calculate DIS Quality Certainty as:

$$C = \Pi_{i=1..n}\, p(Y_i = 1) \qquad (7)$$

This is valid because Y_i are statistically independent random variables (each source quality value varies independently from each other).

In the case of **accuracy** the set of sources $S_1, \ldots, S_n$ are operational when they satisfy one of the restriction vectors $\{v_1, \ldots, v_m\}$ of the accepted configurations. Here, we consider a new the random experiment (over the whole set of sources), where we define:

Random Experiment: verification of the set of sources quality values.

Sample Space: set of all possible combinations of sources quality values.

In this experiment we consider the *event*: “satisfaction of restriction vector v_i of the accepted configurations”, which we note ev_i and is a subset of the random space. The probability of satisfying at least one restriction vector by the sources is the DIS Quality Certainty.

We calculate DIS Quality Certainty as the probability of the union of the ev_i events, which are not disjoint sets:

$$\begin{gathered} C = P(ev_1 \cup ev_2 \cup \ldots \cup ev_m) = \\ \sum_{1\leq i\leq m} P(ev_i) - \sum_{1\leq i<j\leq m} P(ev_i \cap ev_j) + \\ \sum_{1\leq i<j<k\leq m} P(ev_i \cap ev_j \cap ev_k) - \ldots + \\ (-1)^{m-1} P(ev_1 \cap ev_2 \cap \ldots \cap ev_m) \end{gathered} \qquad (8)$$

The probability of one restriction vector is calculated the same way as the DIS Quality

Certainty for one restriction vector (presented for freshness).

In our extended version, we demonstrate that the intersection of two restriction vectors is a restriction vector, by construction. This allows obtaining the probabilities of the intersections that appear in the formula.

3.3 Probability Distribution of DIS Quality

This model gives the probabilities of the different quality values that the DIS may provide in the data targets. For simplicity we calculate it for only one data target, considering the involved sources.

The mechanism for obtaining this model consists on calculating DIS quality values that result from all the possible combinations of sources quality values, and the probabilities of satisfying them. This can be done if and only if the set of possible quality values in each source is finite.

We define *quality-values vector* as a combination of sources quality values:

$vv = <qv_{S1}, \ldots, qv_{Sn}>$, where qv_{Si} is the quality value associated to source S_i

For each DIS quality value it may exist various quality-values vectors that correspond to it.

For the calculation of the probability of each DIS quality value we calculate the probability that one of the corresponding quality-values vectors is satisfied by the sources. For this, we sum the probabilities of the quality-values vectors, since they constitute disjoint events (considering the same random experiment as the one for accuracy in last section).

The following are the steps to calculate the probability distribution of DIS quality, for target T and set of sources $\{S_1, \ldots, S_n\}$:

1- Generate all the possible quality-values vectors for $S_1, \ldots, S_n$, obtaining a set $V = \{vv_1, \ldots, vv_k\}$, where $vv_i = \{q_{i1}, \ldots, q_{in}\}$

2- For each $vv_i \in V$, calculate the quality value provided in T, qv_i, obtaining *DISValues1* = $\{qv_1, \ldots, qv_k\}$.

3- Eliminate duplicate values from *DISValues1*, obtaining *DISValues2* = $\{qv_{i1}, \ldots, qv_{im}\}$, $1 \leq ij \leq k$.

4- For each $qv_{ij} \in$ *DISValues2*, sum the probabilities of the vectors of V that generate qv_{ij}, obtaining the probability that one of the vectors is satisfied by the sources.

Note: The probability of a quality-values vector $vv=<qv_{S1}, \ldots, qv_{Sn}>$ is calculated as:

$P(vv) = P(qv_{S1})\ldots P(qv_{Sn})$, where $P(qv_{Si})$ is given by the source model.

4 CONCLUSIONS

This work proposes an approach to maintain quality on DIS and to deal with source quality changes. To achieve this, we propose to build and maintain probabilistic quality behaviour models of the sources and the DIS, so that they can be used as a support for DIS quality changes detection and management.

The paper focuses on the presentation of the quality models and the techniques for constructing them. A source quality model gives the probability distribution of the possible quality values at the source. DIS quality models give the probability of satisfying quality requirements, the probability distribution of DIS quality values and the satisfaction of quality requirements such as “average”, and “most frequent value”.

The main contribution of the work is the proposal of a novel approach for quality maintenance in DIS, based on quality behaviour models. We believe that the here proposed probabilistic-based approach is an step forward in building quality change-tolerant DIS.

We have worked on the whole mechanism of quality maintenance, which basically consists on relevant quality changes detection and DIS quality repair, but it is not presented here for space reasons.

We have done some experimentation applying our proposal to social networks domain, constructing all the proposed quality models. It showed the usefulness and feasibility of the proposal.

REFERENCES

Canavos, G., 1984. *Applied Probability and Statistical Methods*. Little, Brown. ISBN: 9780316127783

Cho, J., Garcia-Molina, H., 2003. *Estimating Frequency of Change*. ACM Transactions on Internet Technology (TOIT), Volume 3, Issue 3 , Pages: 256 – 290.

Gertsbakh, I., 1989. *Statistical Reliability Theory*. Pub.: M. Dekker. ISBN: 0-8247-8019-1

Scannapieco, M., Missier, P., Batini, C., 2005. *Data Quality at a Glance*. Datenbank-Spektrum 14: 6-1