Managing Source Quality Changes in a Data Integration System

Adriana Marotta

Instituto de Computación, Universidad de la República, Montevideo, Uruguay
amarotta@fing.edu.uy

Abstract. This thesis addresses the problem of source quality changes in Data Integration Systems. Its main goal is to propose techniques for maintaining as much as possible the satisfaction of users' quality requirements. It proposes an approach that bases on a proactive strategy, where probabilistic techniques are applied. These techniques allow to model source quality behavior, and to calculate quality reliability of the system. On the other hand, it proposes a reactive strategy that must be applied for compensating source quality changes. Its detailed solutions focus on freshness and accuracy quality properties.

1 Introduction

Data Integration Systems (DIS) integrate information from a set of heterogeneous and autonomous information sources and provide this information to the users through a mediator schema. These systems basically consist of: (a) a set of autonomous sources, (b) a Mediator, whose data may be materialized or virtual, and (c) the definition of a transformation process, which is applied to information extracted from the sources.

We consider a system where quality properties are taken into account [1]. *Actual values* of these properties are associated to the sources, and *required values* are associated to the mediator for each quality property. Figure 1 shows the system. From the combination of the actual and required values, important design decisions can be made for the improvement of the quality of the DIS (e.g.: source selection).

Some recent works have focused on the evaluation of quality properties in integration systems, i.e. the calculation of the quality values offered by the system to the user [2][3], and also in our research work we have studied this problem [1][4]. We believe there is still much work to do in this direction. However, in these systems, regarding the possible large quantity of sources and their autonomy, another problem emerges that also deserves our attention: sources' quality changes. Actual values of the source elements can change very frequently and in a non-predictable manner, and in general do not evolve going in certain direction, as we can imagine when considering evolution in schemas. We illustrate the importance of this problem with the following example.

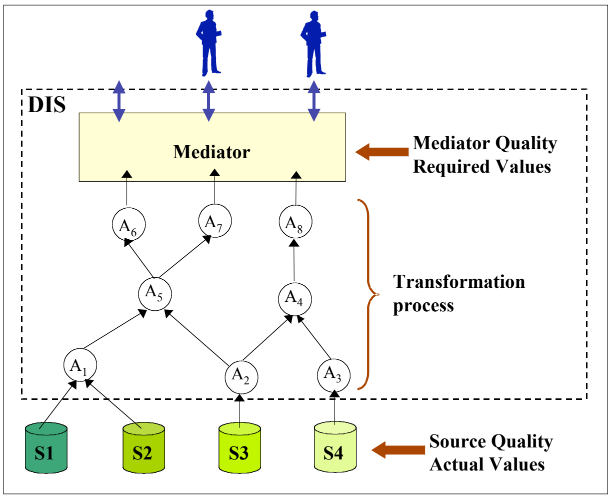

Fig. 1. System architecture

Consider a system that integrates information from several hospitals of a country, which is used by an epidemiology department of the government in order to make decisions concerning the country habitants. The freshness of the information obtained from the system is very important, since the decisions may be quite different according, for example, to the number of cases of certain disease appeared the last day. Therefore, the system was designed so that the users' freshness requirements were satisfied. The system is working as expected when suddenly one of the sources is not updated with the same frequency as before and the information it provides starts being obsolete. In this situation, the system administrator must realize this change, evaluate if it continues satisfying users' requirements and if not, re-accommodate it according to the needs. By this moment, perhaps a decision is waiting to be made or worse, an erroneous decision was already made.

The problem addressed by this thesis is how to manage source quality changes in a DIS. The goal of this thesis is to propose techniques for managing these changes, achieving the users' quality requirements satisfaction as much as possible, and minimizing the impact of the changes over the system. The quality properties focused are freshness and accuracy, not discarding generalizations for other quality properties when possible.

The rest of this paper is organized as follows. Section 2 presents the approach of the thesis, section 3 presents the results obtained up to now, section 4 comments related work and section 5 presents the conclusions and future work.

2 Approach of the Thesis

We identify two strategies to deal with the problem of source quality changes: the proactive strategy and the reactive strategy.

The proactive strategy consists in the actions we can perform before the occurrence of a change that affects the system quality. It includes: (a) techniques for calculating how the quality values of each source can vary without failing to satisfy the quality requirements of the DIS, and (b) techniques based on probabilistic models, which allow to predict source changes and evaluate the system quality reliability. This strategy avoids a lot of useless work in the DIS maintenance, for instance it may avoid source changes, as well as filter cases of source changes that do not affect the quality-requirements satisfaction.

The reactive strategy focuses on the actions we can take after a source quality change occurs, in order to compensate the system quality. It may involve modifications in other quality values of the system without affecting its design, or, when this is not enough to recuperate the satisfaction of the quality requirements, changes on the transformation and/or on the mediator schema of the system.

The thesis has a probabilistic approach that allows to model the behavior of the quality properties at the sources, allows the DIS users to state flexible quality requirements (using probabilities), and provides tools, such as reliability, mathematical expectation, etc., that helps to decide which source quality changes are relevant to the DIS quality. This approach leads to apply a proactive strategy in a continuous basis, and only when necessary a reactive strategy.

3 Preliminary Results

We have worked in three directions: the characterization of the phenomenon of source quality changes [5], the proposal of techniques for the reactive strategy [6], and the proposal of techniques for the proactive strategy. Due to space restrictions we only present the third one.

3.1 Proactive Strategy

The application of this strategy consists in solving two sub-problems: (i) the calculation of admitted quality values for the sources, deduced from mediator quality requirements (Accepted Configurations), and (ii) the modeling of sources quality behavior and the management of changes, basing on probabilistic techniques.

Accepted Configurations.
Given a quality property and a set of sources of a DIS, the *accepted configurations* is a set of combinations of property values for the sources (each property value corresponding to one source) such that the mediator quality requirements are satisfied. Having these possible combinations is very useful, because for example, it is not necessary to re-calculate the system quality each time there is a source quality change, and it may be used as an assistance for negotiation with sources (if you want to compensate a source change asking another source to change some quality value).

Definitions:

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

$$r = qp(R) \text{ op } n, \quad \text{where op may be } <, \leq, >, \geq, \text{ or } = \tag{1}$$

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

$$v = <r_1, \ldots, r_n> \tag{2}$$

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

$$vs = \{v_1, \ldots, v_m\} \tag{3}$$

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

In the case of freshness the restriction-vector space always consists of an only one restriction-vector, so there is a fixed restriction for each source relation. In the case of accuracy, in general a vector space is obtained. The first case is easier and simpler to manage, but more restrictive for the sources quality values. The study of these two cases and some investigation of other quality properties leads us to think that there are two types of properties, those where the source values are independent between each other, i.e. a change in one source does not affect the required value of another source, and those where source values affect each other.

Solutions based on Probabilistic Techniques

We believe the use of probabilities is extremely beneficial for the management of quality properties in this kind of system. It gives more expressiveness to the users at the moment of stating their quality requirements, at the same time making requirements over sources more flexible, it allows modeling source quality behavior, and calculating the system reliability ([7]) with respect to the quality it offers. The use of all these advantages derives in the possibility of better identifying, anticipating and treating the relevant source-quality changes.

We associate probabilities to the mediator required quality values, and to the sources and system quality values. At the mediator, the user has the possibility of stating, attached to each quality requirement, a probability he accepts for its satisfaction. At the sources, we build probabilistic models to model the behavior of the quality values, and we add to the required quality values the probability accepted for them. Globally to the system, we can associate a reliability value, which tells the probability that the system quality requirements be satisfied.

In the case of freshness, for example, suppose we have the random variables X_1, X_2, …, X_n, so that each one correspond to one of the n sources of the integration system. X_i represents the freshness value of source i at a given instant. The restriction vector $v=<r_1,\ldots,r_n>$ (as seen in last section) is the set of conditions that the sources must verify in order to satisfy users' requirements, where $r_i = \text{freshness}_i \leq k_i$. The probability that source i verify restriction r_i is: $p_i = p(X_i \leq k_i)$. The DIS Reliability for freshness property is calculated: $R = p_1.p_2.\ldots.p_n$. Figure 2 shows an example.

Having proposed this meta-data for our system and seen it is useful, in order to define what a source quality change is and how to manage it, we need to work in more specific scenarios, where certain properties of the context are clearly defined.

We consider the quality property freshness and we define a set of possible scenarios based on four dimensions that are relevant to our study: (1) type of DIS, which refers to the materialization or not of the integrated schema (2) type of sources loading (periodically or continuously), (3) meta-data provided by the sources (with respect to updates), and (4) use of the sources freshness values (for calculations or estimations of DIS quality). The crossing of the dimensions generates the scenarios.

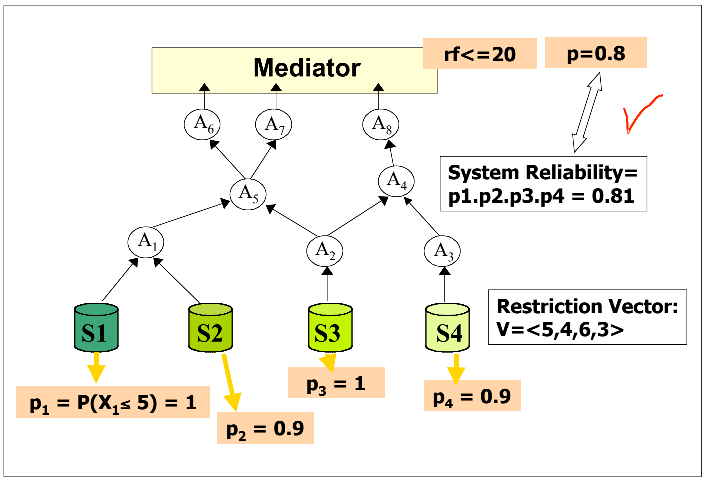

Fig. 2. Freshness and Probabilities

Probabilistic Models

For obtaining some of the previously defined fav it is necessary to model the behavior of the freshness property at the sources. We build probabilistic models where the random experiment consists on the verification of a source freshness value, the random variable (RV) represents the source freshness, and the sample space is the set of all its possible values. The model varies according to the scenario considered. In the following we comment two probabilistic models we have developed.

Model 1:

This model corresponds to a scenario where the sources loading is periodic and the available meta-data is the update period.

Consider a source S, let T be the period of the source loading, and X be the RV for the source freshness. The probabilities for the different possible values of freshness are all equal to 1/T. We also calculate the expectation for X, in **(5)**, which coincides with the average of the possible freshness values.

$$\begin{gathered} p(x) \equiv p(X=x) \\ p(0) = p(1) = \ldots = p(T-1) = 1/T \end{gathered} \qquad \textbf{(4)}$$

$$E(X) = \Sigma_x\, xp(x) = 0(1/T) + 1(1/T) + \ldots + (T-1)(1/T) = (0+1+\ldots+T-1) / T \qquad \textbf{(5)}$$

Model 2:
In this model the sources loading is in a continuous basis. We assume that the source updates follow a Poisson distribution and the source update frequency λ is available.

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 must 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 this time interval:

$$p_U = 1 - p(X=0) \tag{6}$$

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:

$$\begin{aligned} &p(Y=0) = p_U \\ &p(Y=1) = p_U.p(X=0) \\ &p(Y=2) = p_U.p(X=0).p(X=0) \\ &\ldots\ldots \end{aligned} \tag{7}$$

In addition, this model is a stochastic process and also a Markov chain [8].

$\{X_i\}$ is a stochastic process, where X_i is the quantity of updates in each time interval. The X_i are independent RVs and identically distributed with known probability distribution, Poisson.

$\{Y_i\}$ is a stochastic process, where Y_i is the freshness at the end of the i interval. The Y_i are dependent RVs and they can be evaluated iteratively through:

$$\begin{aligned} &Y_{t+1} = \begin{cases} Y_t + 1, & \text{if } X_{t+1} = 0 \\ 0, & \text{if } X_{t+1} \geq 1 \end{cases} \\ &Y_0 = 0 \end{aligned} \tag{8}$$

The $\{Y_i\}$ stochastic process has the markovian property, which roughly means that given the present, the future is conditionally independent of the past:

$$p\{Y_{t+1} = j \mid Y_0 = k_0, \ldots, Y_{t-1} = k_{t-1}, Y_t = I\} = p\{Y_{t+1} = j \mid Y_t = i\} \tag{9}$$

Additionally, the transition probabilities $p\{Y_{t+1} = j \mid Y_t = i\}$ are stationary, since

$$p_{ij} = p\{Y_{t+1} = j \mid Y_t = i\} = p\{Y_1 = j \mid Y_0 = i\} \quad \text{for all } t = 0, 1, \ldots \tag{10}$$

This model gives us the possibility of calculating the transition matrix, obtaining all the transition probabilities, i.e. the probabilities that the freshness changes from certain value to another.

What is a source freshness change?
Considering source-freshness changes has sense when we are doing estimations, since estimations stay constant while its input information does not change. On the contrary, doing calculations, none of the possible source-freshness changes affect us, since each calculation only considers the source-freshness at a given instant. The

source freshness is changing continuously, since the simple time passing causes changes in these values. These are not the changes we want to consider.

For determining the relevant changes, we analyze three sets of changes: (i) those that may affect the DIS freshness, (ii) those that effectively affect the DIS freshness, and finally (iii) those that should generate actions for correcting the DIS freshness. Note that (iii) is included in (ii) and (ii) in (i).

The changes of (i) are those which change our estimations at the source, for instance, in the case of a Poisson distribution, the changes on the parameter λ., and in the case of periodic loading, the changes on the period T. Among these changes only some of them change the DIS freshness. The changes of (iii) are the ones we are searching for, the ones that causes that the DIS freshness requirements are not satisfied. For identifying them we must take into account the freshness requirements stated by the user over the mediator.

Combining the tools previously presented; the accepted configurations, the calculation of the system reliability and the probabilities associated to the freshness requirements, we are able to identify the (iii) set of changes.

4 Related Work

As far as we know, the problem of changes in quality values in data integration systems has not been explicitly addressed. Near to this area, we found the work in [9], which provides a taxonomy of schema evolution operations and the quality properties that are affected by each of them. Nevertheless, we find that the approach in [10] has much in common with ours, since they claim, for example, that instead on measuring accuracy we should focus on change. Their proposal works with baseline values for attributes, detecting statistically significant changes from the baselines. At the same time, we find works that despite not addressing the same problem as we, they serve as a reference point for us, in particular in the application of probabilistic models. For example, in [11] they study how to estimate the change frequency of an element. They assume that a source element changes by a Poisson process, in particular they mention experimental data that shows this behavior for web pages. They focus on estimating λ for different scenarios they characterize by a taxonomy, and they present as further work the problem of changing λ. In [12], they use queue theory for ETL activities, assuming that tuple arrivals to the ETL activities occur due to a Poisson process. Finally, in [13] probabilistic models are used for solving uncertainty of the database values, since they may not coincide exactly with the changing reality. It is also interesting for us, the inclusion to the queries of a probability requirement in the "where" condition.

5 Conclusions and Future Work

The little amount of previous existing work and the problem's wideness and complexity, have leaded us to use an incremental methodology. We are working basically in two iterations: one that overlooks the sub-problems and solutions at a high level, and the second that defines the problems in specific scenarios and proposes detailed solutions for them. This methodology also implies that the detailed solutions are for only two particular quality properties.

The definition of what a source quality change is and how it can be managed is very dependent on: the quality property, the user quality requirements, and the scenario of the DIS (Section 3.3). We have worked on these three aspects so that we have the tools for defining specific solutions for specific quality changes problems.

We think our contribution up to now consists on: (i) identification and characterization of a problem that has not been very much studied before, (ii) study of freshness and accuracy change situations and possible solutions, and (iii) application of probabilistic techniques for the modeling of source quality behavior.

We believe that up to now, the major defect of our work is the lacking of an experimental application, since the proposals where only tested in small study cases. Future work includes working in specific scenarios with the accuracy property, making probabilistic models for this property, and also experimentation of the whole proposal in a real case.

References

1. Marotta, A.; Ruggia, R.: Quality Management in Muti-Source Information Systems. II Workshop de Bases de Datos, Jornadas Chilenas de Computación (JCC'03), Chile. Nov.03
2. Naumann, F.; Leser, U.; Freytag, J.C.: Quality-driven Integration of Heterogenous Information Systems. VLDB 1999: 447-458
3. Bouzeghoub, M.; Peralta, V.: A Framework for Analysis of Data Freshness. 1st Int. Workshop on Information Quality in Information Systems (IQIS'2004). France, June 2004
4. Marotta, A.; Piedrabuena, F.; Abelló A.: Managing Quality Properties in a ROLAP Environment. Accepted paper in 18th. Conference on Advanced Information Systems Engineering. CAISE'06. June 2006. Luxembourg.
5. Marotta, A.; Ruggia, R.: Manejo de cambios en la calidad de las fuentes en sistemas de integración de datos. (content in English). Tech. Report. INCO RT 05-10. ISSN 0797-6410.
6. Marotta, A.; Ruggia, R.: Managing Source Quality Changes in Data Integration Systems. Second International Workshop on Data and Information Quality (DIQ'05) (with CAISE). June, 14th. 2005, Porto, Portugal
7. Gertsbakh, I.: Statistical Reliability Theory. Pub.: M. Dekker, 1989. ISBN: 0-8247-8019-1
8. Hillier, F.; Lieberman, G.: Introducción a la Investigación de Operaciones. Mc.Graw-Hill. 1991. ISBN 968-422-993-3
9. Quix, C.: Repository Support for Data Warehouse Evolution. DMDW'99.
10. Bugajski, J.; Grossman, R.; Tang, Z.: An Event Based Framework for Improving Information Quality That Integrates Baseline Models, Causal Models and Formal Reference Models. International Workshop on Information Quality in Information Systems (IQIS), June 2005, Baltimore, Maryland, USA (SIGMOD 2005).
11. Cho, J.; Garcia-Molina, H.: Estimating Frequency of Change. ACM Transactions on Internet Technology (TOIT), Volume 3, Issue 3 (August 2003), Pages: 256 – 290.
12. Karakasidis, A.; Vassiliadis, P.; Pitoura, E.: ETL Queues for Active Data Warehousing. Int. Workshop on Information Quality in Information Systems (IQIS), June 2005, USA.

13.Cheng , R.; Singh, S; Prabhakar, S.: U-DBMS: a database system for managing constantly-evolving data. Demo. 31st Int. Conference on Very Large Data Bases VLDB '05. Trondheim, Norway, August-September, 2005.