Computación y Sistemas
ISSN: 1405-5546
computacion-y-sistemas@cic.ipn.mx
Instituto Politécnico Nacional
México

Angeles, Maria del Pilar; MacKinnon, Lachlan Mhor
Assessing Data Quality of Integrated Data by Quality Aggregation of its Ancestors
Computación y Sistemas, vol. 13, núm. 3, marzo, 2010, pp. 331-344
Instituto Politécnico Nacional
Distrito Federal, México

Assessing Data Quality of Integrated Data by Quality Aggregation of its Ancestors

Evaluación de Calidad de Datos Integrados por Agregación de Calidad de sus Ancestros

Maria del Pilar Angeles [1] and Lachlan Mhor MacKinnon [2]

[1] Facultad de Ingeniería, División de Ingeniería Eléctrica, Departamento Computación,UNAM.
Edificio "Bernardo Quintana" 2do. Piso, C.U., C.P., 04510 México D.F. Tel. 56223012

[2] Computing & Creative Technologies, University of Abertay Dundee
Dundee DD1 HG, Tel. 01382308601

pilar@macs.hw.ac.uk[1]
l.mackinnon@abertay.ac.uk[2]

Article received on April 29, 2008; accepted on January 05, 2009

Resumen

La calidad de los datos se degrada durante el proceso de extracción y fusión de datos a partir de múltiples fuentes de datos heterogéneas. Además, los usuarios no tienen información acerca de la calidad de los datos que accesan. Este documento presenta los métodos utilizados para la evaluación de la calidad de datos a múltiples niveles de granularidad, incluyendo datos derivados no atómicos teniendo en cuenta la proveniencia de los datos. El prototipo del Manejador de Calidad de Datos ha sido implementado para poder probar dicha evaluación.

Palabras clave: Calidad de Datos, Datos derivados, Integración de Datos, Evaluación de Calidad de Datos.

Abstract

Data Quality is degraded during the process of extracting and merging data from multiple heterogeneous sources. Besides, users have no information regarding the quality of the accessed data.

This document presents the methods utilized to assess data quality at multiple levels of granularity, including derived non-atomic data, considering data provenance. The Data Quality Manager prototype has been implemented and tested to prove such assessment.

Keywords: Data Quality, Derived Data, Data Integration, Assessment of Data Quality.

1 Introduction

Nowadays, users have no information by which to judge data quality. Therefore, the presumption has to be that data is for instance, complete, accurate and current. Most of the existing database systems are based on this "Presumption of Perfection". This is inappropriate because we know that not all data in a database are necessarily perfect, "*Real world data is dirty"* (Hernandez, 1998).

Given the fact of the explosion of online databases created dynamically it is highly unlikely that in any data source all the data would have been primarily authored for that database or that all the values in there are necessarily atomic. These presumptions, the "Presumption of Primary Authorship", and the "Presumption of Atomicity" exist because we have no other way of dealing with data coming from data sources in the existing situation.

Typically, different systems provide conflicting answers to the same question. Extensional data inconsistencies are derived from the integration of independent, distributed data sources.

We present the Assessment Model implementation that provides data consumers with qualitative information of data sources, and derived data within heterogeneous multi-data source environment. Therefore, such qualitative information is given at multiple levels of granularity. This qualitative information could help users when facing extensional data inconsistencies.

2 The Data Quality Manager

The Data Quality Manager has proposed generic and expressive models for the identification and measurement of a set of quality properties. Figure 1 shows the architecture of the Data Quality Manager and the relation between its components.

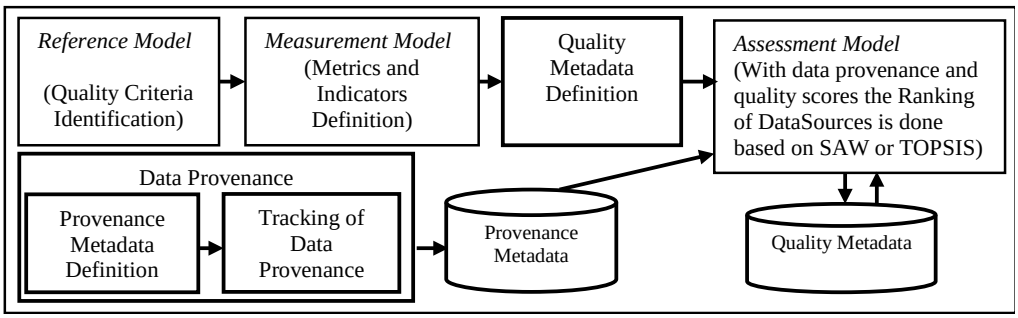

Fig. 1. Architecture of the DQM

2.1 Reference Model

The Data Quality Reference Model contains a set of data quality properties classified and summarized according to different user perspectives such as internal and external focuses or representation, value, and context. Refer to (Angeles, 2004) for further detail.

2.2 Measurement Model

The Data Quality Measurement Model assembles already existing data quality metrics e.g. (Ballou, 1998; Motro, 1998; Pipino, 2002; Naumann, 2003 and Angeles, 2005b). In this document we extend these metrics for the measurement at database, relation, tuple and attribute levels of granularity. Within the Measurement Model, there is an identification of data value level quality properties such as accuracy, consistency, currency, response time, timeliness, uniqueness and volatility. From our research we can state that there are no other systems that use combinations of several quality properties at different levels of granularity as we are using with the Data Quality Measurement Model.

The strictness of quality assessment is a weak or strong characterization depending on evaluating the quality property as a percentage or as a Boolean function respectively as shown in (Scannapieco 2004). The strong characterization of the quality metrics is useful in applications in which it is not possible to admit errors at the corresponding level of granularity.

In the case of the assessment provided by the DQM we have considered the weak strictness to make possible the comparison of data sources for a number of data quality properties. However, there might be alternatives where strictness could depend on the level of quality required, according to specific applications.

In order to assess data quality at different levels of granularity, we have utilised the measures provided at lower levels of granularity (data value, attribute) to determine aggregated scores (table, database) as we move through the levels of granularity. A number of representative metrics for data value, attribute, tuple, relation and database levels of granularity are shown in this section.

Notation: Consider a set *Dm* of domains where each attribute a_1 drawing its values from d_1, a_2 from d_2,… a_m from d_m. A relation *S* is a finite subset of the Cartesian product of one or more domains $d_1 X\ d_2 X \ldots X d_m$ with *m* attributes. The relation *S* is abbreviated as $S(a_1{:}d_1, a_2{:}d_2 \ldots a_m{:}d_m)$. Each element of S has the form $d_1,d_2,\ldots,d_m$ and is called tuple t of relation S. The degree *m* of the relation *S* is the number of attributes in it. The cardinality *n* of the relation *S* is the number of tuples in it. The number of relations in a database D is denoted by *w*.

2.2.1 Consistency

Data value consistency is the extent to which the values for overlapping entities and attributes are the same. Data is consistent with respect to a set of constraints if they satisfy all constraints in the set.

A value in an attribute a_i in a tuple *t* is consistent if and only if it obeys the corresponding constraints.

$$\begin{aligned} Cn_t^{a_i} &= 1 \quad if \quad the \quad value \quad in \quad a_i \quad is \quad consistent \\ Cn_t^{a_i} &= 0 \quad otherwise \end{aligned} \tag{1}$$

The weak consistency at the tuple level $Cn_w(t)$ is the number of instances of the attributes that are consistent divided by the degree of the relation.

$$Cn_w(t) = \frac{\sum_{i=1}^{m} Cn_t^{a_i}}{m} \tag{2}$$

A tuple has strong consistency $Cn_s(t)$ if and only if all attribute instances are consistent.

$$\begin{aligned} Cn_s(t) &= 1 \quad if \quad Cn_t^{a_i} = 1 \quad \forall i \in [1..m] \\ Cn_s(t) &= 0 \quad otherwise \end{aligned} \tag{3}$$

The weak consistency at the attribute level $Cn_w(a_i)$ is the number of tuples where the instance of an attribute a_i is consistent divided by the cardinality of the relation.

$$Cn_w(a_i) = \frac{\sum_{j=1}^{n} Cn_{t_j}^{a_i}}{n} \tag{4}$$

An attribute in a relation is strong consistent $Cn_s(a_i)$ if and only if all the instances of attributes *i* in the relation are consistent.

$$\begin{aligned} Cn_s(a_i) &= 1 \quad if \quad Cn_{t_j}^{a_i} = 1 \quad \forall j \in [1..n] \\ Cn_s(a_i) &= 0 \quad otherwise \end{aligned} \tag{5}$$

The weak consistency at the relation level $Cn_w(S)$ is the percentage of tuples with all instances of the attributes consistent.

$$Cn_w(S) = \frac{\sum_{j=1}^{n} Cn_s(t_j)}{n} \tag{6}$$

A relation has strong consistency $Cn_s(S)$ if and only if all its tuples contain just consistent instances of attributes.

$$\begin{aligned} &Cn_s(S) = 1 \quad if \quad Cn(t_j) = 1 \quad \forall j \quad \in \quad [1..n] \\ &Cn_s(S) = 0 \quad otherwise \end{aligned} \tag{7}$$

Weak consistency at database level is the average of all the consistencies at relation level across the database.

$$Cn(D) = \frac{\sum_{k=1}^{w} Cn(S_k)}{w} \tag{8}$$

The following time related metrics, are considered at tuple level as the direct level of granularity where the measure could be obtained. The level of granularity of the estimation will depend on the DBMS.

2.2.2 Currency

The currency according to (Bovee, 2001; Wang, 1993) is the time interval between latest update and time it is used.

$$Cu(t) = \text{Time Request - last update time} \tag{9}$$

2.2.3 Response Time

Is the delay between the user request and the reception of the complete response from the Information System.

$$RT(t) = \text{Time Reception - Time Request} \tag{10}$$

2.2.4 Volatility

Volatility is the interval of time where data remains valid on the system, and it is related to the update frequency (Ballou, 1998). This dimension characterizes data according with the Information System. For instance, Data Warehouse systems have a very low volatility or no volatility at all, and Transactional systems contain very volatile data.

$$Vo(t) = \text{Update frequency} \tag{11}$$

2.2.5 Timeliness

Timeliness is according to (Gertz, 1998), the degree to which the recorded data are up-to-date. This measure involves not only the currency of data but also if data is in time for a specific usage, and is given under the following terms.

$$T(t) = \max\left(0, 1 - \frac{Cu(t)}{Vo(t)}\right) \tag{12}$$

2.3 Assessment Model

As this is a data-value level approach, each quality property has been assessed through query processes such as parsing, sampling, or continuous assessment as indicated in (Naumann, 2000), and is represented in a quantitative manner. As our purpose is to establish the relative quality among data sources, an exact measure is not necessary.

The assessment of derived data is not possible using the current methods. This led us to the consideration of the process of obtaining the ancestors of a data source namely Data Provenance (DP) in order to estimate the derived data quality from the data sources it has been obtained or computed.

Once obtained the quality scores of data quality at database, relation, tuple, and attribute levels of primary data sources, the following step was the implementation of the data provenance algorithm.

2.3.1 Data Provenance

Due to the huge amount of heterogeneous data sources available to the user, we can no longer rely on the presumptions of perfection, primary authorship and atomicity. Therefore, as our approach is concerned with the quality of integrated data, we have included data provenance as a relevant mechanism to consider in the analysis of data quality.

Woodruff et al and Cui et al have already approached tracking provenance at relational databases (Woodruff, 1997; Cui, 2000).

The algorithm queries the provenance metadata recursively by a lazy approach. This approach is not considering versions or history of data but obtaining enough information to compare one data source against other. The provenance algorithm is not particularly new. It was implemented to obtain enough information for assessment purposes (Angeles, 2005). By considering data provenance, we present two alternatives for assessing derived data presented in the following two subsections.

2.3.2 Assessment of derived data based on the quality of its provenance only.

The provenance algorithm obtains the elements required for deriving data, such as database, relation, attributes and the query utilised. However, such algorithm does not trace the query, so it is not able to analyse the specific conflict resolution function or the formula applied to aggregate data. As a consequence, in the case of an attribute derived from a fusion of attributes, the description of provenance against the elements of interest is required and possible. Therefore, users are able to compare data sources, and to trust one data source against another by comparing the quality properties of their ancestors.

2.3.3 Assessment of derived data by the aggregation of quality of its provenance

Once obtained the ancestors by data provenance, their corresponding quality scores are retrieved to estimate the quality scores of the derived data. As we have no information of the specific resolution functions were utilised for the data fusion, the functions utilised for aggregation of scores are commonly average, maximum, and minimum (Naumann, 2000). The appropriateness of an aggregation function will depend on the optimistic, conservative, or pessimistic approach taken according with the application context. It is not our intension to identify the best aggregation function, because there is not an absolute value. As long as the aggregation function reflects the user needs and it is consistently used, it should be enough for the estimation of quality and comparison purposes.

In this case, the DQM is able to assign quality scores to derived data by the aggregation of the quality properties of its ancestors. This assessment requires that all the quality scores of the corresponding ancestors are available. We are using the aggregation methods average, and maximum.

We have considered average as a conservative aggregation function for accuracy, completeness, consistency, and uniqueness because we require just an approximation of the quality score. An example is explained as follows:

We are concerned with the accuracy estimation of a data source A. Such data source A is a product of data fusion among three data sources with the following scores of accuracy (X=0.65, Y=0.40, Z=0.20). On the one hand, if we take an optimistic approach then the score would be equal to the maximum value, because it is a positive property. Therefore, A is 65% accurate. On the other hand, if we take the pessimist approach, the accuracy score of A is 0.20. However, both approaches are not considering other implicit elements that might change the score. In the former case, A is also fused from data 0.45 less accurate than the maximum value. In the latter case, the score of A has been decreased from the actual value it may have. The accuracy score of A is not 0.65 neither 0.20, but a mix of the three of them. Therefore, fairer estimation would be 0.41 of accuracy, which is the average aggregation function.

From our research perspective, we believe that average as aggregation function is appropriate for comparison purposes. Furthermore, the same approach should be taken on similar situations on positive data value level quality

properties such as consistency or completeness. As another example, in the case of time related properties, we could take a pessimistic approach to have an idea of the oldest data value. Otherwise, users would consider data more current than it actually is.

As the main intention of these approximations is to compare data sources, the option has been to take the most current of the data sources fused. Therefore, the maximum function is used for aggregating scores.

In order to explain how quality of derived data might be assessed through data provenance, consider a query or a source *s* that comes from *n* ancestors α_j

Accuracy of derived data $A(s)$ is computed by the average of the scores of its ancestors.

Completeness of derived data $C(s)$ is determined by the average value of the completeness of its ancestors.

Consistency of derived data $Cn(s)$ is determined by the average of the consistency of its ancestors. The consistency of its foreign keys is checked with its corresponding primary keys in each ancestor.

The currency of derived data $Cu(s)$ is the greatest value of the corresponding currency measures from the different ancestors.

Volatility is the update frequency. When there are a number of data sources with different volatilities, the volatility of derived data $Vo(s)$ is the greatest value of the corresponding volatility measure from its different ancestors

Uniqueness of derived data $U(s)$ is obtained from the average of its ancestor's uniqueness.

Timeliness of derived data $T(s)$ is estimated in terms of its maximum currency and volatility.

$$T(s) = \max\left(0,1-\frac{Cu(\alpha_j)}{Vo(\alpha_j)}\right) \quad , \quad \forall j \quad \in [1\cdots n] \tag{13}$$

We have considered data provenance as a mechanism to help data quality assessment and consequently, to help in the resolution of data inconsistencies between successor databases. We have proposed the extraction of data provenance using a metadata shared by the community, to demonstrate that provenance is as a helpful mechanism to support the determination of data quality. Furthermore, it is possible to trust data according to the quality scores of its ancestors, or to compute the quality of derived data, considering the scores of its ancestors.

Quality of data derived from different data sources at different levels of granularity using data provenance has not been addressed until now.

The Data Quality Manager through the Assessment Model can determine the scores of each participant data source at different levels of granularity. Such scores are stored in the Data Quality Metadata.

Once obtained the quality scores for each participant data source, considering the issue of selecting data sources based on customer priorities in order to obtain the best outcome, it is necessary to take into account the ranking of those data sources based on their quality scores. The ranking of data sources is explained in the following section.

2.3.4 Ranking of Data Sources

On the bases of a comparative established by Naumann (2000), we have utilised two Multi-attribute Decision Making (MADM) methods identified in Hwang (1995) for the ranking of data sources called Simple Additive Weighting (SAW) and Technique for Order Preference by Similarity to Ideal Solution (TOPSIS). These MADM methods have already been used to rank data sources and queries in Naumann (2002), information items in Burgess, (2003), and heterogeneous networks in Zhang (2004).

In the case of positive and negative criteria are involved in the decision matrix we use either Vector normalization scaling method with TOPSIS or Linear scale transformation method with SAW. If there are just positive criteria in the decision matrix, we use either combination. In the case of just negative criteria in the decision matrix, we use either linear scale transformation method with SAW or Vector normalization scaling method with TOPSIS.

We recommend linear transformation function in the case there are only positive criteria or only negative criteria involved, because this scaling method maintains the proportional differences, and such proportion would make a difference in the outcome.

The TOPSIS method is more sensitive to weighting and more sensitive to quality criteria with high scores than the SAW method, which provides a conservative ranking as shown in (Naumann, 2002 and Zhang, 2004).

If users have wide experience, the TOPSIS method should be used for a better performance and sensitivity to users' preferences. On the contrary, if users are naive they should use the SAW method in order to obtain a conservative ranking.

3 The DQM Prototype

We have designed and implemented the Data Quality Manager prototype with an appropriate level of functionality necessary to carry out the proof of concept of this research.

We designed and implemented a repository called Quality Metadata (QMD) to maintain and to retrieve the quality properties identified in the Data Quality Reference Model. The implementation of a repository named Provenance Metadata (PM) was required to store the information related to the data sources involved in the multi-database environment, including the data provenance stored in those data sources.

Analysis of data quality properties

The DQM allows users the analysis of the data quality environment, by identifying which possible cases exist within a multi-database environment and the elements it contains.

We have already mentioned that the DQM is a metadata-based system. Therefore, it is possible that for any reason data information may be incomplete or inaccessible. As the information comes from the Provenance Metadata and the Quality Metadata, we have identified 3 conditions where the DQM can still providing qualitative information:

1. **Analysis of data sources based on their data quality properties only**. There is no ancestor data information at a database level. Therefore, there is no data provenance capability. In such case, The DQM considers all data in the data source as if they were the primary source.
2. **Analysis of derived data based on the quality properties of its ancestors**. There is metadata stored of the data source and that is sufficient to compute quality scores at a certain level of granularity. In the case that the DQM cannot compute quality properties for derived data, the DQM facilitates users the retrieval of data quality by the ranking and analysis of the quality properties of its corresponding ancestors.
3. **Analysis of derived data based on its quality properties**. The quality properties are computed based not on the idea that data are not the primary source but that we have provenance metadata stored at a data source level that describes the ancestor information of the derived data.

Main Functionalities

The main menu is a tool bar representing each element of the architecture of the Data Quality Manager. Some elements of the tool bar are available depending on the status of the user and therefore for a fully working system a password entry system would precede the stage to ensure that user status was identify. Each element is explained as follows:

The Reference Model menu allows the Data Quality Administrator (DQA) the insertion, deletion and update of the quality properties already identified. See Figure 2.

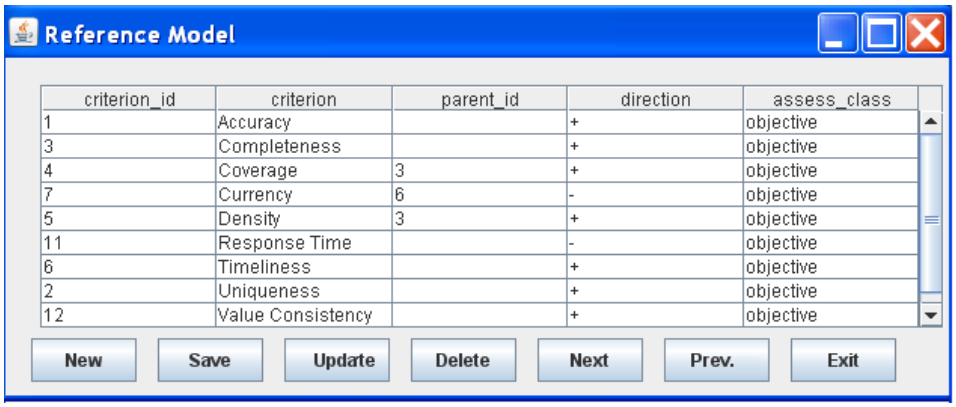

Fig. 2. Maintenance of Reference Model. The Metadata option allows the retrieval and management of the information related to the participant data sources. This activity is only available to the Data Quality Administrator. See Figure 3

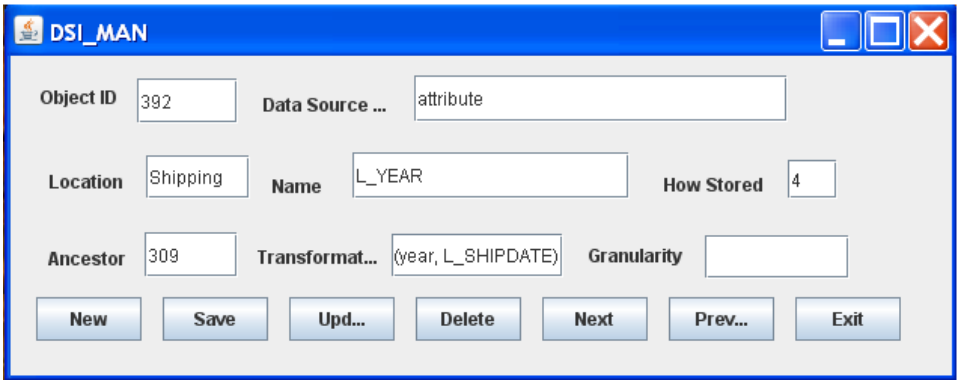

Fig. 3. Maintenance of Metadata. The Measurement Model option contains two submenus Compute Scores and View Scores: The purpose of the former option is to let the Data Quality Manager to compute all the scores within the federation automatically to a certain point to a predefined tolerance period agreed by data consumers. The purpose of the latter is allowing users to retrieve the existing quality scores for a specified data source

The prototype presents users the Assessment Model with a friendly windows interface where they can specify an appropriate context for the analysis of data quality utilising the options: a) Selection of quality properties; b) Specification of quality priorities; c) Specification of data sources to analyse; d) Specification of scaling and ranking methods. Users do not have to select the above-mentioned characteristics individually.

There is a default stereotype condition where scaling and ranking methods are suggested. However, experienced users may wish to select and specify all of the conditions for the context of the query. The system displays a tree with all available quality criteria from the quality metadata for selection, and for each specified quality criterion, a slider is displayed to set the corresponding weight. See Figure 4.

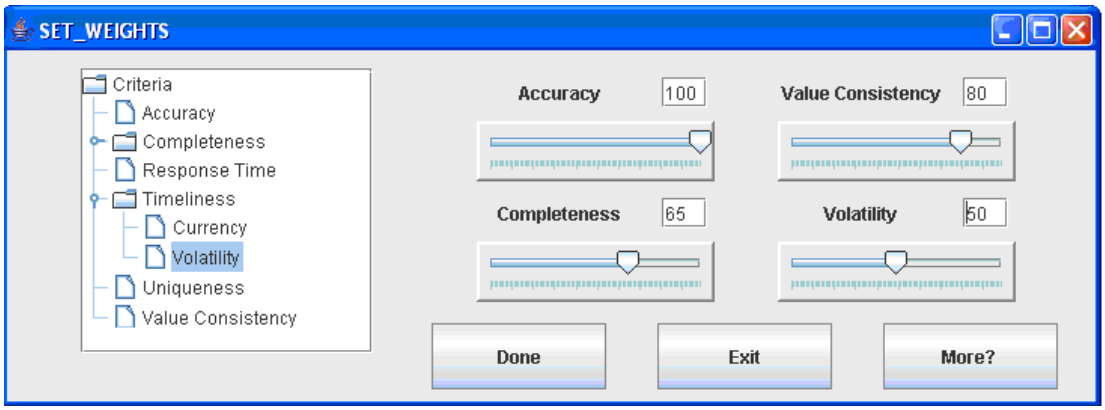

Fig. 4. Prioritisation of quality properties and its priorities

Having all the quality properties prioritised, the weights are normalized. The next step is the selection of the data sources, scaling, and ranking methods. In order to select data sources from a scroll pane, the prototype retrieves from the metadata all the data sources involved in the federation of interest. The scaling method is selected by pressing its corresponding radio button and the ranking of data sources is executed by pressing the buttons TOPSIS or SAW. In the example the linear scale transformation with SAW methods were executed. The overall quality is presented in descendent order in a Text area. See Figure 5.

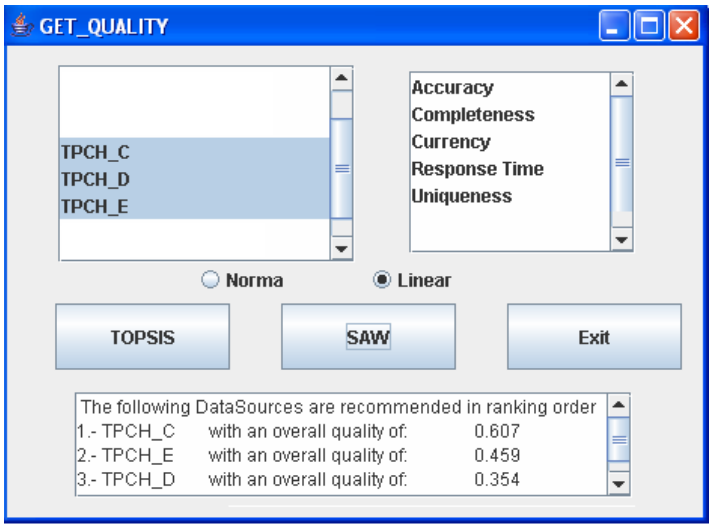

Fig. 5. Selection of Data Sources, Scaling and Ranking methods

The Data Provenance facility obtains the provenance of a selected data source and the quality scores available. Therefore, users are able to select data sources and their quality properties, and then proceed to the ranking of data sources. The first step is to specify the name or identifier of the data source whose provenance is required. The Data Quality Manager provides users with a friendly hierarchical tree describing the data lineage, to trace back to sources through data paths. Every element of the tree can be selected and its available quality scores are displayed in a table. See Figure 6.

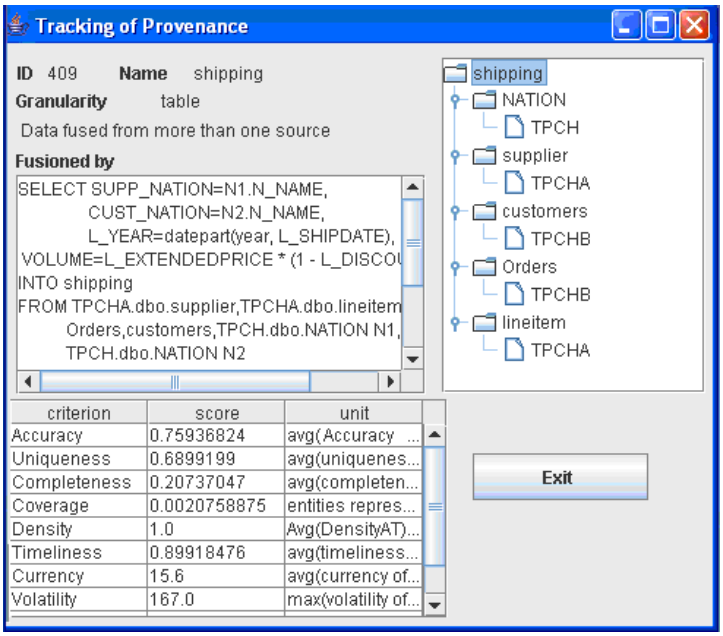

Fig. 6. Data Provenance and available scores of the selected object

4 Testing and Experimentation

We have validated the DQM prototype against the specification of the model, and we have verified that the Data Quality Manager (DQM) can provide appropriate information about the qualitative nature of the data been returned from the data sources.

The performed tests and experiments were on a number of populations corresponding to the following benchmarks:

- The TPC Benchmark™H (TPC-H) for Decision Support Systems, see (TPC-H) for further detail.
- The TPC Benchmark™C (TPC-C) for Online Transaction Processing Systems, see (TPC-C) for further detail.

Testing the prototype

The testing plan considered a number of test cases for testing the functionality of the prototype; testing the appropriateness of the quality information; testing the ranking of the data sources; and testing the appropriateness of the prototype within the test boundaries.

Regarding appropriateness of the qualitative information, we set up the test suite from the TPC benchmarks. We deliberately manipulated the data in some data sources in order to corroborate the intended deficiencies as expected outcomes with the quality scores obtained from the DQM.

We present in this section as a brief example, the test of appropriateness of the qualitative information for a relation called Partsupp from the TPCH benchmark. Table 1 shows the expected percentages of quality and the scores of the Partsupp relation obtained from the DQM prototype.

Table 1. Expected and produced outcomes for relation partsupp

%	Accuracy	Completeness	Consistency	Uniqueness
Expected	90-95%	10-15%	98-100%	12-15%
Obtained	0.9489	0.0119	1.0	0.1446

In the case of the accuracy property, we had estimated a range of 90% to 95% of accurate tuples, and the value produced by the prototype was 0.948, which was within the expected outcome.
For completeness, we estimated a value within the range of 10% to 15% of complete values and the prototype did produce a value of 0.119, which has fallen within the expected range.

As we can see, in the case of consistency we estimated a range of 98% to 100% of consistent tuples. The produced outcome of 1.0 was within the expected range.

Regarding uniqueness, we estimated a range of 12% to 15% of unique values, the outcome presented an outcome of 0.11 producing a value within the range.

As the scores presented by the prototype correspond to the expected outcomes, we could conclude that the DQM prototype performs as expected in terms of the accuracy of the scores obtained. However, this is a limited indicative measurement of the appropriateness of the qualitative information of the DQM. Therefore, an extensive, rigorous testing is required as part of future work. The results of these tests show that the prototype provides appropriate scores according with the expected outcomes based on the actual quality of data.

We designed a set of tests to validate the ranking of the data sources given by the prototype according to their quality scores. The scaling and ranking methods depend on the positive or negative nature of the quality criteria involved in the test. Therefore, we classified the tests as follows: a) Testing the ranking of data sources based on positive criteria only; b) Testing the ranking of data sources based on negative criteria; c) Testing the ranking of data sources based on positive and negative criteria.

In the case of positive criteria, the data source with best score would be the one with the greatest value. In the case of negative quality criteria, the data source with the lowest score value would be the best option. From the measures obtained, we could conclude which data source is the best option in terms of a specific data quality criterion.

In order to test if the qualitative information produced by the DQM varies according to the context, we designed a set of sanity checks in terms of quality criteria, quality priorities, levels of granularity and data provenance through an experimentation plan.

Experimental Hypotheses

The outcome that the DQM produces is that based on a set of criteria, and a query identified by the users, it provides qualitative information about the data sources involved in the query. Such qualitative information can be at a variety of levels of granularity, based on simple information from the data source itself; or based on the information stored in the metadata; or it can explicitly follow a provenance trail to find the original data sources. Hence, ranking of data sources simulations were conducted to identify and prove with high statistical significance (Wilcoxon, 1964; Sheskin, 2004) the following experimental hypotheses:

1. The ranking of data sources changes according to the specification of quality criteria.
2. The ranking of data sources changes according to the quality criteria priorities specified.
3. The query outcome changes according to the assessment at different levels of granularity of their correspondent data sources or in other terms, the ranking of data sources changes according to the level of granularity assessed.
4. A query outcome varies with respect to either one of the following decision criteria: quality properties, assessment with data provenance, and no consideration of quality properties.

A fuller description of testing and the experiments conducted to validate this research can be found in (Angeles, 2007), where there are further details of the test cases, experiments settings, and methods used to measure statistical significance.

5 Contributions to Research

The Assessment Model contributes with the following novel elements:

The Data Quality Assessment Model provides a mechanism for tracking data provenance for the assessment of quality of derived data. Previous approaches work from the presumption of primary authorship and the presumption

of atomicity. Therefore, the introduction of data provenance as a mechanism of considering qualitative information of derived data is novel.

The assessment of derived data based on the quality properties of its ancestors: With the description of provenance, users are able to trace back the quality properties of any data by selecting each data ancestor, having enough information to trust or not to trust the data. In summary, users are able to compare data sources, and to trust one data source against another by comparing the quality properties of their ancestors.

The assessment of derived data based on the aggregation of the quality properties of its ancestors: In this case, the DQM is able to assign quality scores to derived data by the aggregation of the quality properties of its ancestors. This assessment requires that all the quality scores of the corresponding ancestors are available.

Conclusions

This research has considered the association of data quality measures with data retrieved during the querying process on the basis of a set of data sources, a set of data quality properties, and their corresponding priorities.

Databases have been considered as perfect, this research assumes that data are of different and variable quality, challenging the presumption of perfection and providing measures of data quality.

Nowadays, we cannot assume that data sources are the original point of the data they are using. Therefore, discarding the presumption of primary authorship, we have used data provenance as a mechanism to deal with the origination of data within data sources to either assure the user that the data has been provided from the primary source or show evidence of the data provenance and associated quality.

Rather than assuming that data is atomic we can now either assure the user the data is atomic or it is a composed data and demonstrate what the atomic values from which it was generated are, and their associated quality.

Data has been fused, replicated, and transformed to resolve intensional inconsistencies, degrading data quality consequently. Therefore, data provenance along with the quality criteria allows the assessment at different levels of granularity, estimates the quality of derived data, and ranks them at the data value level.

The prototype provides appropriate scores according with the expected outcomes based on the actual quality of data.

We can conclude that it has been possible to identify usable data quality criteria to measure, and assess data quality of primary data sources and derived data at multiple levels of granularity. Such quality information was enhanced by the use of provenance, and the qualitative measures could be used to derive ranking of data sources based on the specification context by the users utilising this known criteria all within a heterogeneous multi-database environment.

Future research

The assessment of derived data based on the aggregation of the quality scores of its corresponding ancestors was illustrative. The analysis of which conflict resolution functions were utilised during data fusion in order to assess data quality with the according aggregation functions corresponds to future research.

The outcome of this work could fit in the future work on Information Quality (IQ) because we have discarded the presumptions of perfection, primary authorship, and atomicity. Therefore, IQ measures could be developed from the Data Quality (DQ) assessment methods we have identified.

References

1 **Angeles P. & MacKinnon L.M. (2004).** Detection and Resolution of Data Inconsistencies, and Data Integration using Data Quality Criteria. *QUATIC 2004: Conference for Quality in Information and Communications Technology,* Porto, Portugal, 87-94.

2 **Angeles, P. & MacKinnon L.M. (2005).** Tracking Data Provenance with a Shared Metadata. *Postgraduate Research Conference in Electronics, Photonics, Communications and Networks, and Computing Science*, Lancaster England, UK, 120-121.

3 **Angeles, P. & MacKinnon L.M. (2005).** Quality Measurement and Assessment Models Including Data Provenance to Grade Data Sources. *International Conference on Computer Science and Information Systems*, Athens, Greece, 101-118.

4 **Ballou, D. & Tayi G. (1998).** Examining Data Quality. *Communications of the ACM*, 41 (2), 54-57.

5 **Bovee, M., Mark, B. & Srivastava E. P. (2001).** A Conceptual Framework and Belief Function Approach to Assessing Overall Information Quality. *International Journal of Intelligent Systems*, 18(1), 51-74.

6 **Burgess, M. S. E. (2003).** *Using Multiple Quality Criteria to Focus Information Search Results*, Ph.D Thesis, Cardiff University, Cardiff, Wales, United Kingdom.

7 **Cui, Y. & Widom, J. (2000).** Practical Lineage Tracing in Data Warehouses. *16th International Conference on Data Engineering (ICDE'00)*, San Diego, California, USA, 367-378.

8 **Hernández, M. A., & Stolfo, S. J. (1998).** Real-world data is dirty: Data cleansing and the merge/purge problem. *Data Mining and Knowledge Discovery*, 2(1), 9–37.

9 **Hwang, C. L. & Yoon, K. (1995).** *Multiple Attribute Decision Making: An Introduction.* London: Sage Publications Inc.

10 **Motro, A. & Rakov, I. (1998).** Estimating the Quality of Databases. *Third International Conference on Flexible Query Answering Systems. Lecture Notes in Computer Science*, 1495, 298-307.

11 **Naumann, F. & Roker, C. (2000).** Assessment Methods for Information Quality Criteria. *International Conference on Information Quality IQ2000*. Cambridge, MA, USA, 148-162.

12 **Naumann, F. (2002).** *Quality-Driven Query Answering for Integrated Information Systems*, *Lecture Notes in Computer Science, 2261*. Berlin: Springer.

13 **Naumann, F, Freytag, J. & Lesser, U. (2004).** Completeness of Information Sources. *Information Systems*, 29(7), 583-615.

14 **Pipino, L. L., Yang, W. L. & Wang, R**. **Y. (2002).** Data Quality Assessment. *Communications of the ACM*. 44(4ve), 211-218.

15 **Scannapieco, M. & Batini, C. (2004).** Completeness in the Relational Model: A Comprehensive Framework. *9th International Conference on Information Quality ICIQ-04*, Cambridge, MA, USA, 333-345.

16 **Sheskin, D. (2004).** *Handbook of Parametric and Nonparametric Statistical Procedures.* London: Chapman & Hall.

17 **Transaction Processing Performance Council (TPC)**. TPC Benchmark C Standard Specification Revision 5.9, 2007.

18 **Transaction Processing Performance Council (TPCH).** TPC Benchmark H (Decision Support). Standard Specification Revision 2.6.2**,** 2008.

19 **Wang, R. Y., Reedy, M. P., & Gupta, A. (1993).** An Object-Oriented Implementation of Quality Data Products. *Workshop on Information Technology Systems*, Orlando, FL, USA. Retrieved from http://web.mit.edu/tdqm/www/tdqmpub/WITS93ObjectDec93.pdf.

20 **Wilcoxon, F. & Wilcoxon, R. A. (1964).** *Some Rapid Approximate Statistical Procedure.* New York: American Cyanamid Co.

21 **Woodruff, A. & Stonebraker, M. (1997).** Supporting fine-grained data lineage in a database visualization environment. *International Conference on Data Engineering ICDE*, Berkeley, California, USA, 91-102.

22 **Zhang, W. (2004).** Handover Decision Using Fuzzy MADM in Heterogeneous Networks. *IEEE Wireless Communications and Networking Conference WCNC 2004*, Atlanta, Georgia, USA, 3-9

Maria del Pilar Angeles *is a PostDoctoral scientist at the Engineering Research Centre of the National University of Mexico (UNAM). She has a PhD in Data Quality from the Heriot Watt University, and a M.Sc. in Computer Science, regarding Quality in Software Engineering from the UNAM. Her research interests are in information quality for heterogeneous databases and quality of software engineering. Since 1989 she has been working as a Technical Support Engineer for Databases at the industry.*

Lachlan MacKinnon*, Head of School of Computing & Creative Technologies of the University of Abertay Dundee. He has a first degree in Computer Science, and a PhD in Intelligent Querying for Heterogeneous Databases. His research interests are in information and knowledge engineering. He has published extensively in this area, worked on and chaired national and international conferences, and led research projects in the UK and Europe. He is a member of the IEEE, British Computer Society, ACM and AACE.*