Quality and Recommendation of Multi-source Data for Assisting Technological Intelligence Applications

Laure Berti

GECT, Systèmes d'Information Multi-Media
University of Toulon, B.P. 132, F-83957 La Garde cedex, FRANCE
berti@univ-tln.fr

Abstract. Due to its costly impact, data quality is becoming an emerging domain of research. Motivated by its stakes and issues, especially in the application domain of Technological Intelligence, we propose a generic methodology for modeling and managing data quality in the context of multiple information sources. Data quality has different categories of quality criteria and their evaluations enable the detection of errors and poor quality data. We introduce the notion of relative data quality when several data describe the same entity in the real world but have contradictory values : *homologous data*. Our approach differs from the general approach for resolving extensional inconsistencies in integration of heterogeneous systems. We cumulatively store *homologous data* and their quality metadata and we recommend dynamically data with the best quality and data which are the most appropriate to a particular user. A value recommendation algorithm is proposed and applied to the Technological Intelligence application domain.

1 Introduction

Accentuated by the Internet phenomenon and the development of Information Technologies, heterogeneity and volume of data flows make economically and humanly difficult the retrieval and the critical analysis of useful information. The user is instantaneously submerged by data which, either do not precisely meet his/her needs, or when relevant, are incomplete, vague and/or contradictory. As a matter of fact, the issues of information searching are moved from the quantitative stakes to the qualitative ones (i.e. from data volume to data quality). Moreover, the validation of information remain difficult. This context requires, not only, a much more active and critical attitude of the user, but also, serious abilities for information interpretation, evaluation, deduction, analysis and synthesis. Although information systems are advantageously used to store, manage and analyze quantitatively information items, few of them propose assistance for examining critically the quality of their information content. When data are stored in the database, they are generally not qualified or certified according to their relative quality and their effective utility. But, it is essential to explicitly evaluate the quality of data stored in databases in order (1) to adapt information quality according to the querying audience, (2) to provide a

T. Bench-Capon, G. Soda, and A M. Tjoa (Eds.): DEXA'99, LNCS 1677, pp. 282-291, 1999.

critical expertise of the quality of the information system content, (3) to enable the user to put into perspective his/her confidence in data and, make him/her change or adapt his/her data usage. Data quality becomes an emerging theme of research and development in various industrial [14, 8], commercial [11], military and scientific [10] application domains. Nowadays, the current approach to measure data quality is statistical by sampling important volumes of data. But, in the profusion of contradictory information, the quality of data can be evaluated by comparison (i.e. by comparing the quality of *homologous data* : data which are extracted from different information sources, which describe the same reality but have contradictory values). Rather than to solve the existing conflicts between the different values of data, our approach is to exploit these conflicts to evaluate the relative quality of data. Because information quality expertise is a necessary daily task in several strategic application domains such as Technological Intelligence, we developed the sQuaL system to assist information quality expertise and to propose an adaptive selection of data with the best quality. Data recommendation is made according to the relative quality of data. As result of a query, the sQuaL system presents the data with the most appropriate quality among the candidate *homologous data*. The user can choose different strategies for recommendation.
The remainder of the article is organized in the following way : Section 2 presents previous works on data quality and introduces our contribution concerning the quality of multi-source data. Section 3 describes a general methodology for integrating the quality expertise of multi-source data in the information processing. Section 4 defines an algorithm for data recommendation. Section 5 illustrates an application example of the sQuaL system in Technological Intelligence. Section 6 concludes and presents our perspectives of research.

2 Data Quality : the multi-source perspective

Data quality aroused an increasing interest since a long time, but, it is now clearly becoming an emerging field of research [13,14,11,10,15]. Globally, the various works on the data quality can be classified in three research avenues according to their objectives : (1) to define each dimension of data quality with a rigorous scientific method (2) to create a universal standard set of operational quality dimensions (3) to let data quality dimensions be defined by users and to propose assistance for data quality evaluation.

Modeling data quality. Since 1980, many propositions have been made for modeling data quality [4,7,14,11,10] but the consensual definition of data quality is still not reached. A very complete analysis [13] presents the vast panorama of research on data quality. Several works completely integrate the modeling and the management of data quality into the design of information systems (by using labels on every element of the conceptual model) [15,12]. The use of metadata for the improvement of data

quality is also recommended in many standards of exchange for Geographical Information Systems (ISO 15046-13, CEN 287-008... [5]).

Evaluating data quality. The current approach for measuring data quality is a statistical approach centered on methods such as the inference on missing data and automatic control of data exceptions. Many methods were developed to measure the four essential data quality dimensions : accuracy, completeness, currentness and consistency [9,7,8] in conformance with users' quality specifications. But actually, data quality audits [13,8] are the only practical means for determining data quality in databases.

Relative data quality. In parallel with the frequencist approaches in data quality research, some subjectivist and user-oriented approaches propose assistance to users. In this perspective, we extend the proposition of [15] by modeling the ***relative quality of data*** which includes three sub-categories of quality criteria [3] : the first sub-category of criteria is related to the context of the application (based on a frame of reference such as time or application), the second sub-category of quality criteria is related to the user (based on the cognitive or emotional frame of reference), and finally, we introduce the multi-source perspective into the characterization of data quality : the quality of one data can be evaluated by comparison with the quality of other *homologous data* : i.e. data from different information sources which represent the same reality but have contradictory values. And we propose recommendation mechanisms based on the relative quality of the *homologous data*. Our problematic joins numerous works on integration of heterogeneous systems, in particular, for what concerns the problems of extensional inconsistencies (problems of tuple identification, conflicts of values...). Although these problems are unanimously admitted, few solutions have been proposed to solve the conflicts between contradictory values of data. Finally, the approaches usually adopted to reconcile heterogeneities between values of data are : (1) to prefer the values of the most reliable sources (2) to mention the source ID for each value (3) to store quality metadata with the data.

3 Methodology for multi-source data quality expertise

In a multi-source information context, users need to be sure about the quality of data they use to make important decisions. This implies the contribution of specialists for examining quality and value-adding data. Here, a particular function of the information system (IS) is to store information quality expertise and to recommend the most appropriate data to the user. Thus, for the same query, different information results may be proposed to various querying users according to their data quality requirements. Consequently, the IS should be able: (1) to cumulate *homologous data* : in the database, the same attribute may have multiple values proposed by different information sources, (2) to assist experts when they cross-check information items

and when they evaluate the relative quality of data. In order to present the application context of the sQuaL system, we propose the 7 steps of the following methodology.

Step 1. Detecting needs

In this preliminary step, the needs must be clearly expressed, in particular, the information quality required by the various users (human operators, technological specialists, decision makers...). Requiring a real implication of these actors, the purpose is to specify the vocabulary of the application domain, as well as the vocabulary of the dimensions (criteria) of the information quality concept and to make sure that every body uses the same concepts with the same meaning. The goal of this step is to build an ontology (as a reusable modeling schema) supporting the modeling of intellectual capital resources of the company (information and information quality expertise).

Step 2. Selecting quality criteria

Directly dependent on the first step, the purpose of this second step is to choose several data quality criteria to be measured/evaluated by human specialists for determining the quality of information sources and the quality of their content (as structured data). Two types of quality criteria are to be measured : (1) objective quality criteria (2) subjective quality criteria (authors reputation, credibility...) evaluated by experts.

Step 3. Selecting information sources, collecting and mapping data

The goal of this step is to select relevant information sources, to collect and store their data into the database and finally, to establish correspondences between *homologous data* which describe different versions of the same reality (*mapping*). The selection of information sources is based on the evaluation of selected quality criteria (Step 2). Data flows must be precisely described : actors who can modify data quality and their interventions on the database (creation, update, use, deletion) must be identified. This step underlines attributes that are incomplete or, complete but never used. It raises the question of the effective utility of attributes and it suggests the formalization of decision rules and the distinction between various quality levels for the attributes according to their particular use. The mapping of *homologous data* is supervised by the specialist, expert of the information field. As a result in the multi-source database, attributes have multiple values with the mention of the ID of their information source (Fig. 1).

Step 4. Selecting critical data

The goal of this step is to select the classes, objects and attributes that are *critical* for the application domain. Every data do not have the same importance, they are not equivalent from an strategic point of view : they should not be considered with a uniform way. The concept of criticality is used to compare the importance of data. The *critical* objects and attributes are classified and selected according to their importance with respect to the application objectives. They compose the ***subset of critical data***.

Step 5. Evaluating information quality

The goal of this step is the information quality expertise and the storage of information quality metadata for each critical object and attribute. Each quality criteria can be measured by : (1) direct measurements and (2) indirect measurements with objective or subjective quality indicators. Procedures for objective measurement and protocols for subjective evaluation must be explicitly described. In the case of conflicts between contradictory values from various information sources, quality evaluation mainly depends on knowledge and competence of the specialist who supervises data capture, evaluates the quality of the data and affects subjective quality indicators to information items. A tolerance level for data non-quality can be expressed according to the criticality degree of the data. Quality metadata are associated with each critical data (Fig. 1) and stored in the ***quality metadatabase***.

Step 6. Identifying non-quality problems and analyzing their causes

The goal of this step is to reveal non-quality problems through the entire information processing (errors or poor quality data) and to analyze their causes. After having identified the actors and the processes which create, maintain, collect, visualize or use each object and attribute, it may appear that a (sub-)process is misused or does not satisfy users. The efforts focus on the identification of problems occurring in the manual or computerized activities. This analysis must identify in particular : who carries out the selection of multi-source data, their mapping, the evaluation of quality ? What is the user satisfaction ?... The purpose of this step is also to search the causes of data non-quality by classifying the events which disturb the information chain.

Step 7. Recommending value-added data and defining new quality objectives

The analysis and the evaluation of data (non-)quality enable recommendation mechanisms. The recommendation must be adaptive, multicriteria and dynamically computed according to each new data arrival and user quality preferences. The company must also lay down objectives for improving its information quality for the

mid and long terms, define resources and strategies to be implemented and ensure the support and the improvement of the quality of its data. The strategy of maintenance of the database and the whole information chain must take into account the evolution of the user's needs in terms of data quality.

4 Multicriteria recommendation based on relative data quality

The objective of the methodology was to provide a global vision for the integration of data quality expertise in a multi-source information environment and also to clearly present the application context of our content-quality-based recommender system : the sQuaL system [2,3]. In this section, we describe the structure of data and quality metadata and the recommendation mechanisms used in the sQuaL system.

4.1 Multi-source objects and quality metadata

In the multi-source database, each attribute of a multi-source object has multiple values with the ID of their source and their associated quality expertise (Fig. 1). Quality expertise is represented as metadata associated with each value. Each quality criterion may have a subjective evaluation indicator chosen by the specialist and justified when it's necessary (Fig. 1 : Justificative Comment). The scale of subjective quality evaluation we chose is : 0=very low - 0.25=low - 0.50=medium - 0.75=high - 1=very high.

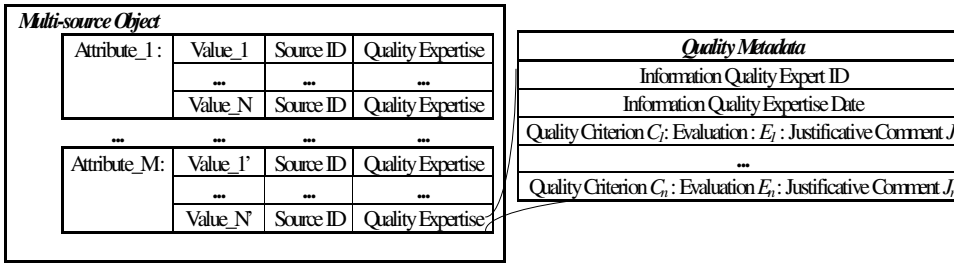

Fig. 1. Multi-source object structure with associated quality metadata.

4.2 Data quality scoring and ranking for recommendation

The objective of the multicriteria recommendation based on the relative quality of data is to propose the *most appropriate* data ; that is, the data which are correct and have the best quality required by the user : for example, he/she may privilege the freshness of the data rather than the credibility of their information source. It is often necessary to take into account the relative importance of the quality criteria and to affect specific weightings on quality criteria for the recommendation. We apply four

existing methods in Operational Research [6,1] for data recommendation. For each data value, a score reflects its relative quality according to one of the following quality scoring model :

– *Linear Assignment of Weight (**LAW**)* which defines a quality score for each data :

$$Qscore(D_i) = \sum_{k=1}^{l} W_k E_{ik} \text{ with } W_k : weight\ of\ the\ k^{th}\ quality\ criterion$$
$$E_{ik} : evaluation\ of\ the\ k^{th}\ quality\ criterion\ for\ the\ i^{th}\ data\ D_i$$

– *Elimination by Aspects (**EA**)* : this method classifies the quality criteria by importance, then eliminates data that have the worst quality score for the most important criterion. The operation is repeated until there remains only one data for the recommendation.

– *the method of Anderson (**AND**)* [1] and *the method of Subramanian and Gershon (**SG**)* : these method uses three matrices (concordance, discordance and magnitude matrices) indicating each one the relative situation of one data compared to its *homologous data*.

4.3 Multicriteria recommendation algorithm

Using the recommendation models, we determine the data with the best quality. For each candidate value in one attribute of one multi-source object, we define a quality score accumulator and propose the following recommendation algorithm :

```
Algorithm Value_Recommendation

Let Vacc[v] be the entry for the value v in the quality
           score accumulator
Init all the accumulator entries to 0
For each multi-source object o
  Find the set of attributes {a1, a2,..., al} of o
  /*l : number of attributes describing the object o*/
  For each attribute ai (with i = 1,..., l)
    Find the set of values {vai,1, vai,2,..., vai,mi} of ai
    /*mi : number of values for the attribute ai
          proposed by the different sources*/
    For each value vai,j (with j = 1,..., mi)
      For each quality criterion ck in the metadata
      associated with the value vai,j (with k = 1,...,nj)
       /*nj : number of quality criteria for vai,j*/
       Find the set of values {vck,1, vck,2,... ,vck,ml} for
       which the criterion ck has been evaluated
       /*ml : number of values for which the quality
              criterion ck has been evaluated (mi≤ml)*/
          For each value vck,p (with p = 1,..., ml)
            Find the source s of the value vck,p
            Vacc[vck,p] = Vacc[vck,p] + Vscore(vck,p, ck, s)
Sort the score accumulator by decreasing order
Return the list of values according to score
accumulator value
```

The Vscore($v_{ck,p}$, c_k, s) is the function which gives the score of the value $v_{ck,p}$ for the quality criterion c_k tacking into account the quality score of its source s. This function

uses one of the scoring methods previously presented, according to the particular recommendation strategy.

5 Application of sQuaL in Technological Intelligence

In the context of Information Warfare, an Information Service Provider, such as a Technological Intelligence Group (*TIG*), is composed of human experts, specialists whom competencies are used to evaluated the veracity of available contradictory information items. Explicit and tacit knowledge about plausibility of technological information are essential knowledge to be capitalized. In this perspective, the aim of the sQuaL system is to assist the tasks of the *TIG* basically in charge of : (1) selecting information sources (technological patents, technical reports...). The selection of producers is based on objective and subjective quality criteria (e.g. credibility, interest to disclose information...) (2) choosing and collecting dynamically information : the selection of presumed relevant information is based on non-exhaustive information quality criteria : plausibility, accuracy, timeliness, completeness... (3) cross-checking information items (contradictory or partially redundant), evaluating and stamping information quality and information source quality, (4) providing certified relevant information.
The general methodology we proposed in Section 3 is particularly appropriate to the Technological Intelligence application domain [2,3] whose main tasks are : (A) collecting the descriptions of real entities made by various information sources and storing relevant and critical data in the multi-source database (B) evaluating data quality criteria (C) recommending data for consultation. Let's now present a short example of application in the context of Technological Intelligence.

A. Selection of sources and collection of data and quality metadata. For instance, a competitor's product ***P*** is a real entity which is never entirely and/or precisely known by the *TIG*. To know its characteristics, the *TIG* doesn't have necessarily a direct access to the equipment ***P***. Thus, the *TIG* must trust relevant textual information sources. In Fig. 2, three textual information sources S1, S2 and S3 propose characteristics to describe the product ***P*** : respectively, "*length : 42f*" for S1, "*length = 40 feet*" for S2, and "*long:15.023m*" for S3. Initially, the TIG experts have to collect all the descriptions made on the product ***P*** by the available information sources. Then, they structure and store these descriptions in the multi-source database of the sQuaL system (Fig. 2). *Homologous data* are stored in *multi-source objects* whose attributes have multiple values with the mention of their source ID. In the example, the product ***P*** has three values for the attribute *length*.

B. Evaluation of quality criteria. In the example, we suppose that the four quality criteria (selected in Step 1 and 2 of the methodology) are : the plausibility of the data value, its accuracy, the credibility of its source and the freshness of the data. These criteria may be ranked per degree of importance by the TIG expert (Plausibility: 60%,

Accuracy: 20%, Source Credibility: 15% and Freshness: 5%). The quality criteria of each value are evaluated by the specialist of the technological field and stored as quality metadata. In the example of Fig. 2, the plausibility of the value "*42 f*" proposed by the source S1 is evaluated *High* by the expert.

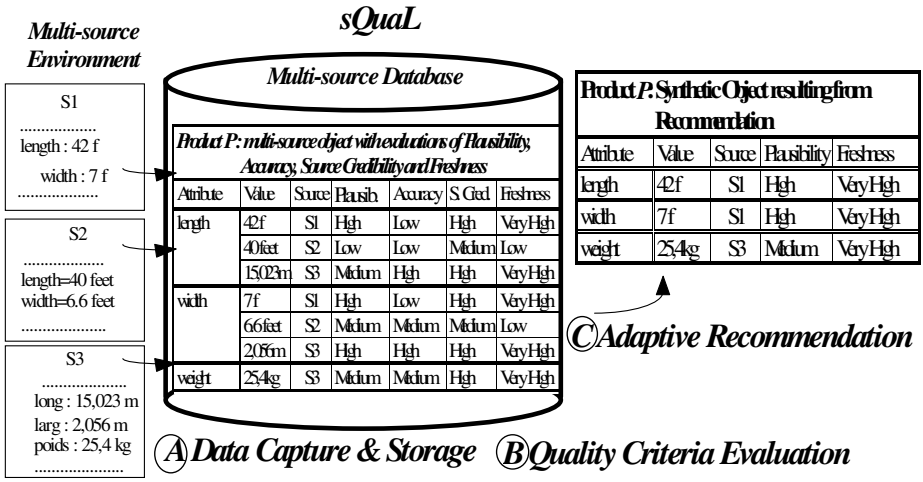

Fig. 2. Application example of sQuaL

C. Adaptive value recommendation. Querying the sQuaL system, the user may have requirements for data quality and may rank the quality criteria he/she considers more important. So, the recommendation results are adapted to his/her data quality needs. The sQuaL system will propose an adaptive value recommendation for each attribute of the multi-source object. For answering the query, a synthetic object is built with the existing values of the multi-source object according to the best quality scores. In the Fig. 2, the user was interested in the plausibility and the freshness of the data (privileging plausibility). A synthetic object is finally proposed to him by the sQuaL system : its attributes are composed of values from heterogeneous sources (S1 and S3) for their optimal plausibility and freshness.

6 Conclusion and perspectives

The objectives of our article are : (1) to briefly present the existing works on data quality and to introduce the notion of multi-source data quality (2) to propose a general methodology for integrating information quality expertise into the information processing (3) to present the sQuaL system for assisting information quality expertise and recommending data according to their relative quality in the specific context of Technological Intelligence. Our current work is now focused on :

(1) the assistance for the choice of the best recommendation strategy (2) the anticipation of the sQuaL system for non-intrusive acquisition of user profiles and quality criteria preferences (3) the fuzzy recommendation. The perspectives of research on multi-source data quality and their recommendation are numerous, in particular for extending of the concept of *relevancy-ranking* with the concept of *quality-ranking* for (semi-)structured data.

References

1. Anderson, E., Choice models for evaluation and selection of software packages, J. of Management Information Systems, Vol. 6, (1990) 123-138
2. Berti, L., Out of overinformation by information filtering and information quality weighting, Proc. of the 2nd Information Quality Conf. MIT (1997) 187-193
3. Berti, L., From data source quality to information quality : the relative dimension, Proc. of the 3rd Information Quality Conf. MIT (1998) 247-263
4. Brodie, M.L., Data quality in information systems, Information and Management, Vol. 3 (1980) 245-258
5. Goodchild, M., Jeansoulin, R., (eds), Data quality in geographic information : from error to uncertainty, Hermès (1998)
6. Fritz, C., Carter, B., A classification and summary of software evaluation and selection methodologies, Technical Report, Mississippi State University (1994)
7. Fox, C., Levitin, A., Redman, T., The notion of data and its quality dimensions, Information Processing and Management, Vol. 30, no. 1 (1994)
8. Redman, T., Data quality for the information age, Artech House, (1996)
9. Reddy, M. P., Wang, R., Estimating data accuracy in a federated database environment, Proc. of the 9th Intl. Conf. CISMOD (1995) 115-134
10. Smith, I., Pipino, L., (eds), Proc. of the 3rd Information Quality Conf. MIT (1998)
11. Strong, D., Kahn, B., (eds), Proc. of the 2nd Information Quality Conf. MIT (1997)
12. Wang, R., Kon, H. B., Madnick, S. E., Data quality requirements analysis and modeling, Proc. of the 9th Int. Conf. on Data Engineering (1993) 670-677
13. Wang, R., Storey, V., Firth, C., A framework for analysis of data quality research, IEEE, TKDE, Vol. 7, no. 4 (1995) 623-638
14. Wang, R., (ed), Proc. of the 1st Information Quality Conf. MIT (1996)
15. Wang, R., A product perspective on Total Data Quality Management, Communications of the ACM, Vol. 41, no. 2 (1998) 58-65