An innovative data quality diagnostic tool in a business intelligence context

Manuel Mejía-Lavalle, Ricardo Sosa Ríos[1]

Instituto de Investigaciones Eléctricas, Reforma 113, 62490 Cuernavaca, Morelos, México, [1]Centro Nacional de Control de Energía, Comisión Federal de Electricidad, México DF, México

mlavalle@iie.org.mx, ricardo.sosa@cfe.gob.mx

(Paper received on September 09, 2010, accepted on October 20, 2010)

Abstract. Nowadays, database technology is a computing area with a great growth. Today is common that we can hear about databases with huge volumes of information and also we can hear about Business Intelligence projects related with these huge databases. However, in general, little attention has been given to the quality of the data. Here we propose and present an innovative software tool designed to perform a basic task related to the Data Quality issue, this is, the diagnosis. Our tool takes ideas from the Artificial Intelligence discipline to realize the diagnosis task in an automatic and in a human-like way. The initial results that we obtained when we apply this tool to a very large and real world database encourage us to continue this research.

1 Introduction

Databases are one of the computing areas with more acceptance and growth, especially since the devices for storing large volumes of data have become very efficient and inexpensive. Not many years ago we can found that a database close to the Gigabyte (10^6 bytes) was already very large. Currently is common that we can hear about databases that store, for example, Terabytes (10^{12} bytes) and Yottabytes (10^{24} bytes) of information, and the trend is increasing.
But this high growth has generally accompanied with little attention to data quality that these databases contain, being now more than ever true the old phrase at the beginning of the computer days that said "garbage in, garbage out". And while there is abundant literature on the subject, there are few concrete proposals for software tools that directly address the issue of data quality for large databases.

Given this problem, in this paper we propose and present a software tool designed to perform one of the key tasks related to data quality, i.e. diagnosis, which involves measuring the level of quality in a database. The proposed tool integrates ideas taken from the Artificial Intelligence (AI) area to perform an automatic and human-like diagnosis. This tool has been tested with a very large database of the electric sector and the results that we obtained have been satisfactory, as described in this article.

For developed these ideas, first we will address the issue of Business Intelligence, diagnosis by means of AI and Data Quality in general, and then we will present the tool that we propose, describing a new measure designed by us and used to establish an objective diagnosis of the data quality; also we will describe how this tool operates, being designed in a generic way to work with various databases and platforms; finally

©G. Arroyo-Figueroa (Ed)
Special Issue in Advances in Artificial Intelligence and Applications
Research in Computing Science 51, 2010, pp. 57-66

we will summarizes the results and we will discuss the conclusions and the work to be performed in the immediate future.

2 Business Intelligence, AI and Data Quality

Nowadays, huge corporations are seeking to know more about their business processes. They usually have enormous and valuable data repositories, but they do not know what to do with these data. It is common to hear the phrase: "worse than have too little (or any) data, is to have many data and not knowing what to do with it" [1].
Business Intelligence (BI) can be a useful approach to meet the challenge. BI is focused on transform data into knowledge (or intelligence) to improve corporation central process. At the end, BI is a discipline formed with tools emerged from AI and Database technology, which main purpose is to give people the information or knowledge that they need to do their jobs.
The term BI was coined by Howard Dresner about twenty years ago [2], to describe an emerging discipline concerned with the discovery of information (that was not known before) in a corporation. BI includes methodologies and tools like:

- Data Warehouses [3],
- On Line Analytical Processing (OLAP) and related methods (MOLAP, ROLAP, etc.) [4],
- Knowledge Discovery in Databases (KDD) and Data Mining [5],
- AI areas and algorithms like, for example, Machine Learning, Intelligent Multi-Agents Systems, Artificial Neural Networks, Fuzzy Logic, Case Base Reasoning, Pattern Recognition, Genetic Algorithms, Expert Systems, etc. [6],
- Statistical analysis,
- And, in general, any algorithm, tool or method that serve to transform data into knowledge.

It is predicted that, in the near future, BI will become a need of all huge corporation [2].
But maybe the first great challenge of BI is manage information that contains data with appropriate quality. Speaking in a broad context Data Quality refers to conduct a thorough investigation of the data in the database. This research can be done before to the creation of the database or for those already in operation. It includes determine who are the users of the database, what they need, what is the essence of the business, what are the important variables, how often the information is required, what level of detail is required, with what levels of safety and risk is needed, etc. And, for those databases in operation, we need measure the current quality of information, in order to known and improve that information.
The activities of defining, measuring, analyzing and improving the data in the database results in the total quality management data cycle, which sees information as a product and is a powerful methodology to develop and maintain databases that contain quality data which is required by the business and is based on the principles of quality proposed by Deming [7]. According to Hufford [8] Data Quality consists of

five basic dimensions: completeness, validity, consistency, timeliness and accuracy, which together mean that the data are appropriate for a particular purpose.

Although Data Quality should be a starting point must for every computer system with databases, in practice this objective is not met in most of the cases. And even with a quality system in place, the experts are agree in the sense that any large database can have a 100% quality, as mentioned by the international computer systems analyst company Gartner [9].

Thus, since we cannot achieve a perfect database that meets all the requirements expressed by the Data Quality theory, a remedy to ensure that a database is useful, initially, is to focus only on the dimension named "accuracy", identifying dirty data and diagnosing the quality of data in order to apply cleaning (data cleansing or data cleaning). This cleaning process can include removing those records or variables that, according to some criterion, are dirty, duplicate or unuseful. Another more sophisticated type of cleaning is by means of estimate statistically the possible value of dirty data based on data believed to be clean, or by inferring it, applying AI ideas [10], [11].

A special form of data with noise is when the data is unknown, and then Kononenko [12] identifies several types: forgotten or lost, not applicable, irrelevant, or omitted in the design. Brazdil [13] has proposed ways of dealing with unknown values, and in particular Quinlan [14] has worked with AI top-down induction of decision trees techniques for the handling of unknown values, and has proposed up to seven different treatment schemes.

An important part of data cleaning is to check the consistency of records, i.e., detect whether there are cases with the same values of attributes (or similar) with different classes [15]. A special case is when the cleaning process is over non-numeric attributes, i.e., they are text descriptions, such as names of people, products, address, etc.: in that case the cleaning has to be developed based on an AI parser program to detect similarities and standardize and verify the data [16].
For the innovative software diagnostic tool proposed here, we have used the concepts of BI, AI, Data Quality and data cleaning described above to identify dirty data and thus obtain a general analysis of the database. These topics are detailed in the next section.

3 Proposed Software Diagnostic Tool

Among the objectives of the tool that we present for the diagnosis (in an AI fashion) of the quality of a database, we can mention the following:

- Obtain an initial approach to the problem,
- Getting a general idea of the status of data (global view – focused on the business data),
- Measuring data quality,

- Establishing patterns of data quality,
- To detect critical points in the data, and
- Being able to have a starting point to develop the cleaning business rules to be applied to the data.

To describe the tool that we developed, first we will discuss the metrics that we devised to obtain a numeric indicator of the quality level of the data, in an objective way. Then we will describe how this tool operates, being designed in a generic way to work with various databases and platforms. Finally we will discuss the results that we obtained by applying this tool to a large database of the electric sector.

3.1 Proposed Metric

There are a number of metrics designed to obtain an indication of the quality of the data. In particular we focused our research work on the dimension "accuracy" of data. We seek for a metric that was simple, so it could be easily understood, yet robust, to be able to get data quality information at different levels of data aggregation, i.e. at the attribute level, the table level or at the database level. Additionally, we seek that our metric can accept a weighted schema (assigning costs depending on the importance of each attribute or table), and we seek that it was supported by the experience of other companies related in the quality issue. We also seek that the metric may include different types of dirty data, from the most common, even those who are less frequent.
Our metric is based on the "Frequency check" that is used by: Cambridge Research Group [17], Knowledge Integrity Incorporated [18], Business Objects (recently acquired by SAP) [19], Group 1 [20] and Gartner [9], all these are solid companies in the Information Technology area.
In our case, we have one error per each incorrect or missing data, and we sum all occurrences and we named like "#incorrect". The accumulated error is expressed as a percentage according to:

$$\% \text{ Error} = \#\text{incorrect} / \text{total data} \quad (1)$$

where "total data" is obtained in various ways, depending on the level of aggregation. For an attribute the variable "total data" is equal to the total number of records; for a table the "total data" value is obtained by multiplying the number of attributes in the table by the number of records; for a database it is calculated by the sum of the "total data" of each table in the database.
In the event that a field has no data, an error is registered. In the case of an attribute with no data, it is assigned a 100% error to this attribute. In the case of a table with no data, also it is assigned a 100% error to this table.
To assign weights to the attributes or important tables, 100 points should be considered for all attributes of a table. Then these 100 points are distributed according to the importance of each attribute (representing the weight assigned for the user). If we have a total of 10 attributes, each would have 10 points if we want that all the attributes had the same weight. Thus, the weight serves as a factor that is applied to each attribute to obtain the value of "%Error" in a weighted schema. In other words,

"% Error" reflects the fact that there are attributes with greater weight than others. The same idea would be applied to the table level.
According to the above expressed, the quality is calculated as:

$$\text{Quality} = 100 - \%\ \text{ERROR} \qquad (2)$$

Then, if "Quality" is 100% we have a perfect database and if "Quality" takes a value of 50% we can say that the database is wrong in a half of its data. The importance of this measure is that it permits to have an objective measure such that it is able to independently evaluate certain attributes of interest for a particular user, or evaluates a single table that is of particular importance, or show, in a comprehensive manner, the quality of a complete database, all this depending on the special information needs of each user.

3.2 Tool Description

The diagnostic tool that we propose allows for an automatic human-like analysis of the data quality of a specific database, through three aggregation levels: a) Attribute, b) Table, and c) Database. The tool based his diagnosis by means of identify missing values (blanks), zero (never caught), repeated characters, dates and numbers out of range, etc. The diagram in Figure 1 shows the relationship among the several windows that are part of the tool: the hierarchy is shown in terms of how each window interacts with the user.

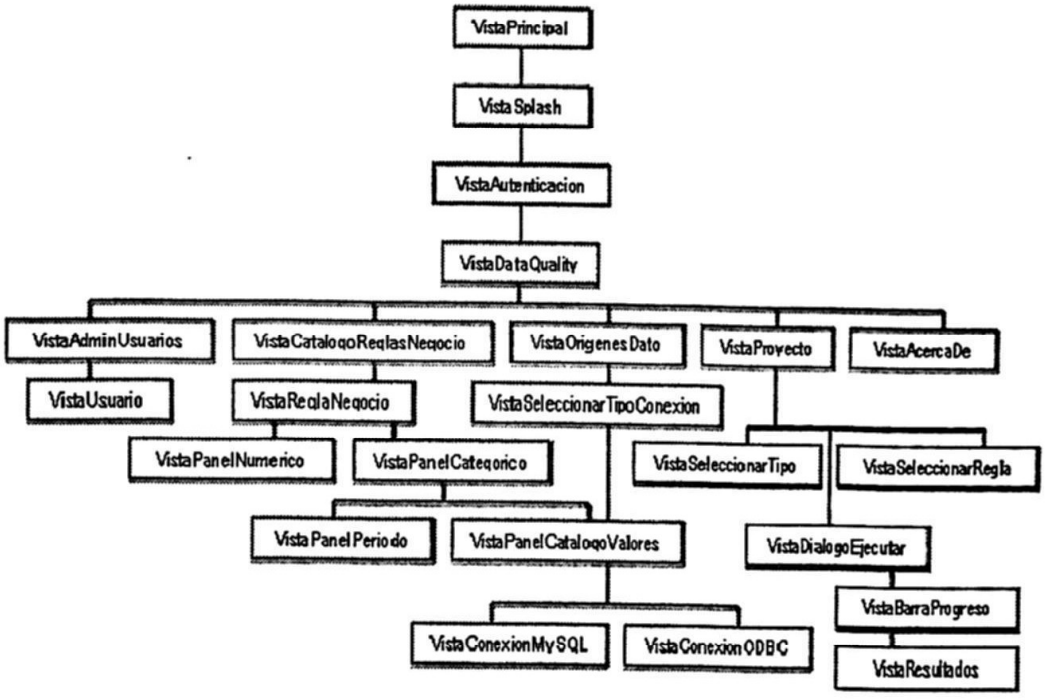

***Fig. 1** Schematic hierarchy windows diagram of the proposed diagnostic tool.*

The central idea to search for and identify bad data is to conduct a count of the number of occurrences of each of the values of an attribute occurs in the table: data that appear very infrequently can be considered as "suspicious", and this basic idea is

applied by us to numerical values and also to text values of an attribute. This idea is detailed paragraphs below.

An innovative feature is that the tool is designed for flexibility, since it has the characteristic of being configurable to access various sources of data (platforms) to create various BI rules that are capable of detecting suspicious quality in data.
The tool has the ability to connect to various data sources by means of JDBC (Java Data Base Connectivity) technology or via ODBC (Open Data Base Connectivity). Additionally, the tool manages business rules and they assist the diagnostic process, serving as indicators to identify incorrect or anomalous values. In our tool it is possible to define a business rule catalog (or knowledge base), which can later be used in different "cases of diagnosis", relating each rule with multiple attributes to support the process of quality data diagnosing.

The business rules are a particular type of production rules, traditionally used in Expert Systems. We design our tool like an Expert System Shell [21] in order to gain several advantages from this area, like: capability to create and increase expert knowledge by means of new production rules, include common sense knowledge, obtain permanent expertise, achieve easy to transfer and document rules, gain consistency, capability to verify knowledge and obtain expertise in an affordable way.

In our tool there are two types of business rules: for text data and numeric data. In the case of text data, the business rule can detect out of range data (only accepts a set of predefined valid descriptions), incorrect data, dates out of range, null data, data with repeated characters and missing data. For numeric data, the tool detects out of range values by grouping into 21 intervals all the data, being the first and last intervals (often with infrequent data) those that can be considered suspect. The basic ideas were taken from the AI area, according to [10]-[16] and from de Data Quality area [9], [17]-[20].

For example, to create a business rule to detect strange symbols, null values and repeated characters, the user just have to select the "Text type" button, followed by the "Special characters" option and click the "Ok" button. To create a business rule to detect values out of range of a numeric attribute, the user only have to select the "Number type" button, then define a valid range and click the "Ok" button. Once defined and stored all the necessary business rules, the user have created a catalog (or knowledge base) of business rules, which may be applied to the attributes which she or he considers necessary and appropriate to link.
The tool also allows to the user to create and store cases of diagnosis: this tool feature allows to the user easily run this pre-defined diagnoses cases, without necessity of rewriting the business rules. This feature is similar to have several domain experts like in Expert System is realized. To do this, the user specifies a title of the event (diagnostic case), the period of data to analyze, the business rules assigned by attribute, and sets the data source, tables and attributes to diagnose. Figure 2 shows a tool interface as an example of how diagnosis cases are introduced using simple graphic elements. We summarize the data quality diagnosis method in Figure 3.

After the execution of a "diagnosis case" the tool automatically generated three types of reports: a) Frequency Values Report: the tool generates an outline of the analyzed data by means of a frequency list of values that each attribute has. If some assigned business rules is related, the report also shows a column with the number of errors found by that rule, b) List of rules applied, and c) List of detail records where errors were detected.

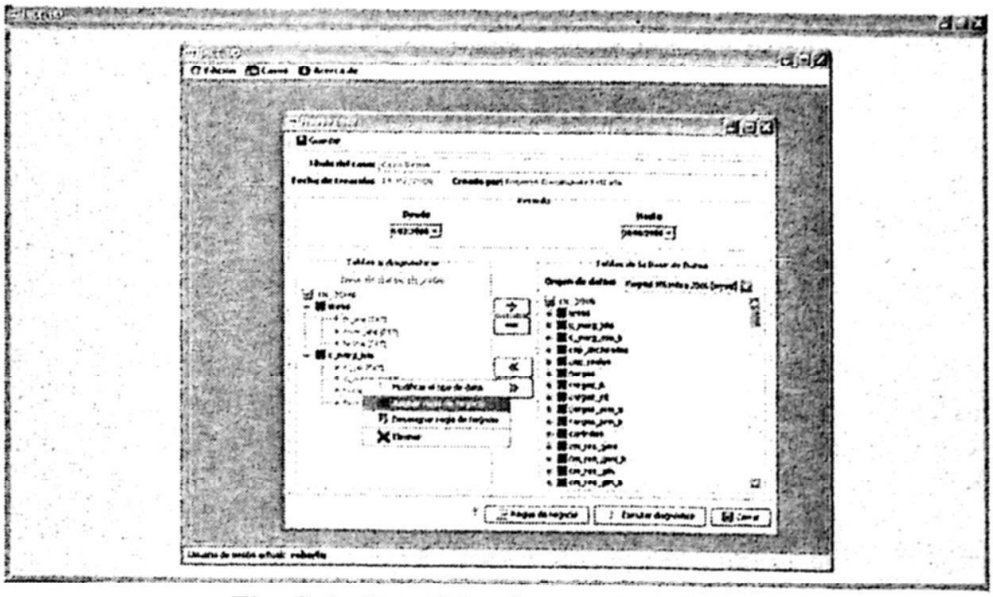

***Fig. 2** A view of the diagnostic tool interface.*

Given a database with M tables, and each table with D attributes and N instances,

1. Initialize variables *%Error*, *Quality*, *#incorrect*;
2. Assign user-expert estimated weights to attributes and tables;
3. For each M, D and N:
   Apply business rules from knowledge base to numeric or text data
   If an error is detected, increment *#incorrect*,
4. Calculate global metrics *%Error*, *Quality* at different aggregation levels;
5. Print quality reports.

***Fig. 3** Summarization of the proposed tool (data quality diagnosis phase).*

At the moment to publish this paper, the data cleaning phase is not applied yet. But the same scheme of business rules for diagnosis can be used for the data cleaning phase. In Figure 4 we show a possible algorithm proposed by us to infer unknown data, following the ideas from [14]. In particular we propose step C2 to use the well known ID3 algorithm applied to the unknown data problem.

3.3 Results

The tool described here was used successfully to analyze and diagnose a large database of the Mexican Federal Electricity Commission (CFE) related with the electricity market. Our tool was capable of analyzed nearly 200 tables containing

more than 2,000 attributes that represents about 2 billion data. With the tool was able to detect whether there were attributes with errors, and if there were some tables more problematic than others. For reasons of confidentiality we do not show details. But in general, we can say that the information obtained using the proposed diagnostic tool is appropriate to improve the quality of the data, like the final users point out during the test period.

A. Find the attribute that better divides the data set into homogeneous subsets: for each attribute, calculate the disorder or entropy according to the following formula:

$$E = \sum_{r} [Nr/Nt] [\sum_{c} \{-(Nrc/Nr) \log_2 (Nrc/Nr)\}]$$

Nr = number of examples in branch r
Nt = total number of examples in all branches
Nrc = total of examples in branch r of class c

B. The attribute which has the smallest value of E is taken as the root node of the tree (attribute-node) and there will be one branch for each value that the attribute has[1].
C. For each value of the attribute-node, select all the examples (rows) with the same attribute value. For each subset do the following:

C1. If all examples belong to the same class, the branch is labeled with the class.

C2. If the subset is empty, find the most similar example (smaller distance) to the current branch; if the distance is acceptable (according to certain threshold previously defined), label the branch with the class of the most similar example, otherwise label the branch as "unknown class".

C3. If the examples in the subset belong to different classes, go to step A, with this subset as the new data set.

D. If there are branches without labels, go to step A, otherwise finish.

Fig. 4 ID3- based algorithm to infer unknown data (data cleansign phase).

In particular, we consider that the results were successful because we can meet the initial project objectives like: a) To obtain an initial approach to the problem: at the beginning we don't know the CFE's databases data quality situation, and after apply the tool we obtain a better idea of the dimensions of the problem and then it could be possible propose to CFE several future action schemes in order to increase database quality, b) To get a general idea of the status of data, detecting in a global view and focused on the business data, the reality of the data, c) To obtain a objective measuring of the data quality, i.e., a qualification or score, that represents a starting point to initiate a total quality management project, d) To establish a group of initial detect critical points in the data that needs immediate attention, and f) To be able to

---

[1] The idea of a complete tree was taken from [22].

have a starting point to develop the cleaning business rules to be applied to the data in order to increase in an automatic and human-like way the quality of the database.

4 Conclusions and Future Work

We present a novel software tool for the diagnosis of the quality of data in large databases, in the context of BI and taken ideas from the AI area. In particular we describe an innovative measure in an objective way to measure this quality on the dimension "accuracy" of the data and able to obtain indices at different levels of data aggregation, i.e. at the attribute level, the table level or at the Database level. The results obtained by applying this AI based tool to a large database of the electric sector have been successful, because the tool was capable to detect wrong data immersed in billions of data. With the data conveniently clean, we can now initiate properly a BI development.

As future work, we see that it would be important to add to the diagnosis tool the ability to create business rules to find dirty data in an inter-relationships among attributes way, i.e. to find when one or more data make that other data be "dirty "because they lack the proper context. To give a simple example, one can consider the case of an attribute or field of "personal names" that could be validated against the attribute of "sex of the person", so this require that the name of the person was appropriate to their gender, otherwise, would be marked as an error or a like a wrong captured data. Additionally, we need aggregate a more complete inference mechanism to the tool, in order to take more advantage from the Expert Systems ideas (i.e., symbolic reasoning) and can manage more sophisticated diagnosis schemas. Also it will be important add an explanation facility to justify how the tool reaches a particular data quality diagnosis.

Acknowledgements

The authors thank and acknowledge the involvement of CFE staff, especially of the CENACE centre, which provided valuable comments to define the design specifications of the diagnostic tool. In particular, authors thanks to the engineers: Nemorio González, Eduardo Roa, Anselmo Sanchez, Pedro Alatorre, Jorge Lazaro and Rikica Romano.

They also recognize the work of development and implementation of MC Roberto Dominguez Estrada and acknowledge the helpful suggestions from the anonymous reviewers to improve the final version of the paper.

References

1. Richeldi, M. (1999). A business intelligence solution for energy budget control. Proceedings of the 3rd International Conference on the Practical Application of Knowledge Discovery and Data Mining, (167-82).
2. McKay, L. (2008). Business intelligence comes out of the back office. CRM magazine, Jun .
3. Gill, H. S. (1996). Data warehousing, Prentice Hall Hispano-americana, S.A.

4. Brackett, M. H. (1996). The data warehouse challenge, John Wiley & Sons, Inc.
5. Piatetsky-Shapiro, G. (1991). Knowledge Discovery in Databases: An Overview, In Knowledge Discovery in Databases, Piatetsky-Shapiro, G. eds., Cambridge, MA, AAAI/MIT.
6. Turban, E., Aronson, J., Liang, T., Sharda, R. (2005). Decision support and business intelligence systems, Prentice Hall.
7. Huang, K., Lee, Y., Wang, R. (1999). Quality information and knowledge. Prentice-Hall, NJ.
8. Hufford, D. (2009). Data warehouse quality, DMReview, www.Dmreview.com/ editorial/dmreview/
9. Gartner, 4th (1999). Annual Enterprise Technologies Summit, Centro Banamex - Ciudad de México, April.
10. Arroyo, G., Villavicencio, A. (1994). Intelligent system to improve heat rate in fossil power plants. In Proc. of the IFAC Artificial intelligence in real-time control, Valencia, Spain, pp. 33-39.
11. Ibarguengoytia, P. (1997). Anytime probabilistic sensor validation. PhD Thesis, ITESM, México.
12. Kononenko, I. (1992). Combining decisions of multiple rules. In Boulay, B. (ed) Artificial Intelligence (AIMSA), Elsevier science Pub, pp. 87-96.
13. Brazdil, P., Bruha, I. (1992). A note on processing missing attribute values. Canadian Conf. on AI, Workshop on Machine Learning, Vancouver, B.C., Canada.
14. Quinlan, J. (1989). Unknown attribute values in ID3. Int. Conf. on Machine learning, pp. 164-168.
15. Bruha, I. (2000). From machine learning to knowledge discovery: survey of preprocessing and postprocessing. Intelligent data analysis, IOS Press 4: 363-374.
16. Kimball, R. (1996). Dealing with dirty data, DBMS on line. www. dbmsmag.com/9609d14.html, Sept.
17. http://research.microsoft.com/en-us/labs/cambridge/ [consulted on January 2010].
18. http://knowledge-integrity.com/wpblog/ [consulted on March 2010].
19. http://www.sap.com/solutions/sapbusinessobjects/index.epx [consulted on February 2010].
20. http://www. G1.com/Support/ [consulted on October 2009].
21. Waterman, D. (1986). A Guide to Expert Systems, Addison-Wesley Publishing Co.
22. Guzmán, A. (1995). Sustitución de Sistemas Expertos por Arboles k-d, Congreso de Reconocimiento de Patrones, CIMAF-95, La Habana, Cuba, January.