Data Quality Evaluation previous to Big Data Analytics

M. Mejia-Lavalle*, J. Perez-Ortega, A. Magadan-Salazar, E. Perez-Luna,
G. Reyes Salgado and D. Mujica Vargas
Departamento de Ciencias Computacionales, Centro Nacional de Investigación y Desarrollo Tecnológico,
Int. Internado Palmira S/N, 62490, Cuernavaca, Morelos, México
{mlavalle*, magadan, eduardo.perez, greyes, dantemv@cenidet.edu.mx}, jpo_cenidet@yahoo.com.mx

Abstract - *Big Data technology is a computing area with a great growth. Today it is common that we hear about databases with huge volumes of information and also we hear about Data Mining and 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 innovative metrics and schema designed to perform a basic task related to the Data Quality issue, this is, the diagnostic. The preliminary results that we obtained when we apply our approaches to Big Data encourage us to continue this research.*

Keywords: Data quality, Big data, Data cleansing.

1 Introduction

Big Data is one of the computing areas more active for research, especially since the devices for storing large volumes of data have become very efficient and inexpensive. Not many years ago we found that a database close to one Gigabyte (10^6 bytes) was considered very large. Currently it 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 been generally accompanied with little attention to the quality of data 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 Big Data subject, there are few concrete proposals for schemas that directly address the issue of data quality for very large databases.

Given this problem, in this paper we propose and present some metrics and a schema designed to perform one of the key tasks related to data quality, i.e. diagnostic, which involves measuring the level of quality in a database. In other words, our approach realizes a data quality evaluation, previous to initiate a Big Data analytics phase and it has been tested in a preliminary project and the results that we obtained have been satisfactory, as described in this article.

For developing these ideas, first we will address the issue of Big Data and data quality in general, and then we will present the ideas that we propose, describing new measures designed by us and used to establish an objective diagnosis of the data quality; also we will describe how these metrics operate joined to a modified machine learning algorithm, being our design developed in a generic way in order to work with various databases and platforms; finally we will summarizes the preliminary results and we will discuss the conclusions and the work to be performed in the immediate future.

2 Data Mining, Big Data and Quality

Nowadays, huge corporations are seeking to know more about their business process. They usually have enormous and valuable data repositories, but they do not know what to do with this 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].

Data Mining, Knowledge Discovery in Databases (KDD), Business Intelligence or Big Data Analytics, can be useful technologies to meet that challenge. These approaches are focused on transforming data into knowledge (or intelligence) to improve corporate central process. At the end, the term Big Data represents a computer discipline formed with tools emerged from Artificial Intelligence and Database technology, which the main purpose is to give people the information or knowledge that they need to do their jobs.

Before Big Data and after Data Mining, the term Business Intelligence (BI) was coined by Howard Dresner several years ago [2], to describe an emerging discipline concerned with the discovery of information (that was not known before) in a corporation. BI includes disciplines 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],
- Artificial Intelligence areas and algorithms like, for example, Machine Learning, Intelligent Multi-Agents Systems, Artificial Neural Networks, Fuzzy Logic, Case Base Reasoning, Pattern Recognition, Genetic Algorithms, 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 corporations [2]. But, more recently the term "Big Data" has emerged. According to [7] Big Data can be characterized by: a) volume (large amounts of data), b) variety (includes different types of data), and c) velocity (constantly accumulating new data).

But maybe the first great challenge for Big Data is to manage information that contains data with the 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 determining 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 will be required, what level of detail is required, what levels of safety and risk is needed, etc. And, for those databases in operation, we need to measure the current quality of information, in order to know 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 [8]. According to Hufford [9] 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 agree in the sense that any large database can have a 100% quality, as mentioned by the international computer systems analyst company Gartner [10]. Thus, since we cannot achieve a perfect database which 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 un-useful. 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 [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 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., there are text descriptions, such as names of people, products, addresses, etc.: in that case the cleaning has to be developed based on a parser program to detect similarities and standardize and verify the data [16].

For the metrics and schema proposed here, we have used concepts from Big Data, BI, KDD, data mining, 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 Data Quality Diagnosis Metrics and Schema

Among the objectives of the schema that we present for the diagnostic of the quality of a very large database, we can include the following:

- Obtain an initial way of how to attack the problem,
- Get a general idea of the status of data (global view – focused on the business data),
- Measure data quality,
- Establish patterns of data quality,
- Detect critical points in the data, and
- Reach a starting point to develop the cleaning business rules to be applied to the data.

To describe the data quality evaluation schema 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 approach operates, being designed in a generic way to work with various databases and platforms. Finally we will discuss the preliminary results that we obtained by applying this schema to simulated large databases.

3.1 Metrics for Data Quality

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 data 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 [10]; all these are solid companies in the Information Technology and Big Data areas.

In our case, we define 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:

$$\%Error = \#incorrect / TD \quad (1)$$

where *TD* stands for "total data" and it is obtained in various ways, depending on the level of aggregation. For an attribute the variable *TD* is equal to the total number of records; for a

table the *TD* 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 relevance than others. The same idea would be applied to the table level.

According to the above expressed, the quality is calculated as:

$$Quality = 100 - \%Error \quad (2)$$

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 shown, in a comprehensive manner, the quality of a complete database, all this depending on the special information needs of each user.

3.2 Data Quality Schema Description

As stated before, our schema allows for an automatic analysis of the data quality of a specific database, through three aggregation levels: a) Attribute, b) Table, and c) Database. Additionally the proposed approach based his diagnosis by means of identifying missing values (blanks), zero (never caught), repeated characters, dates and numbers out of range, etc. There are relationships among the several characteristic data blocks: the hierarchy is established in terms of how each characteristic data block interacts.

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 that occurs in the table: data that appear very infrequently can be considered as "suspicious dirty", 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 our design seeks for flexibility, since it has the characteristic of being configurable to access various sources of data (platforms) to create various Business Intelligence rules that are capable of detecting suspicious quality in data.

This design 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 proposed schema manages business rules and they assist the diagnostic process, serving as indicators to identify incorrect or anomalous values. In our shema it is possible to define a business rule catalog, which can later be used in different "cases of diagnosis", relating each rule with multiple attributes to support the quality data diagnosing process.

The business rules are a particular type of production rules, traditionally used in Expert Systems. We design our schema 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.

We define 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 schema detects out of range values by grouping into a predefined quantity of intervals, being the first and last intervals (often with infrequent data) those that can be considered dirty-suspect.

For example, to create a business rule to detect strange symbols, null values and repeated characters, the user just has 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 has 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 has created a catalog of business rules, which may be applied to the attributes which she or he considers necessary and appropriate to link.

The schema also allows the user to create and store cases of diagnosis: this feature allows the user to easily run this pre-defined diagnoses cases, without necessity of rewriting the business rules. 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.

After the execution of a "diagnosis case" the schema automatically generated three types of reports:

a) Frequency Values Report: is 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 suspicious data or errors were detected.

We summarize the data quality diagnosis algorithm in Figure 1.

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 the 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. 1 Summarization of the proposed schema (data quality diagnostic phase).

At the moment to write 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 2 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 Diagnosis results for simulated Big Data

The schema described here was used successfully to analyze and diagnose a large academic database. Our schema was capable of analyzed nearly 200 tables containing more than 2,000 attributes that represents about 2 billion data. With the prototype was able to detect whether there were attributes with errors, and if there were some tables more problematic than others. In general, we can say that the information obtained using the proposed diagnostic schema is appropriate to improve the quality of the data, like candidate users point out during the test period.

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 databases data quality situation, and after apply the prototype we obtain a better idea of the dimensions of the problem and then it could be possible propose 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 patterns of data quality, that can be enriched with the time instead to be lost, e) To detect critical points in the data that needs immediate attention, and f) To be able to 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.

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:

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

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.

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

4 Conclusions and Future Work

We present a novel software schema for the diagnostic of the quality of data in large databases, in the context Big Data. 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 schema to a large academic database have been successful, because the prototype was capable to detect wrong data immersed in billions of data. With the data conveniently clean, we can now initiate Big Data analytics properly.

As future work, we see that it would be important add to the diagnosis schema 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 prototype, 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 schema reaches a particular data quality diagnostic.

5 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] Berman, J., (2013). Principles of Big Data, Elsevier Inc.

[8] Huang, K., Lee, Y., Wang, R. (1999) *Quality information and knowledge*. Prentice-Hall, NJ.

[9] Hufford, D., *Data warehouse quality*, DMReview, www. Dmreview.com/editorial/dmreview/

[10] Gartner, 4th Annual Enterprise Technologies Summit, Centro Banamex - Ciudad de México, Abril 1999.

[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 2016].

[18] http://knowledge-integrity.com/wpblog/ [consulted on March 2016].

[19] http://www.sap.com/solutions/sapbusinessobjects/index.epx [consulted on February 2016].

[20] http://www. G1.com/Support/ [consulted on October 2015].

[21] Waterman, D. (1986). A Guide to Expert Systems, Addison-Wesley Publishing Co.