



Identification of Data Quality Issues in Homogeneous Environment

Sapna Pujara¹, Dr. Kanwal Garg²

¹Research Scholar, ²Assistant Professor

^{1,2}Department of Computer Science & Applications, Kurukshetra University,
Kurukshetra, Haryana, INDIA

Abstract: Maintaining the data quality of the database simply means passing the data through a standard set of test and correct procedures before it may be utilized for any information retrieval. The database may be piled up with the data, collected from different kind of sources or it may be collected from geographically- distributed sites but homogenous in respect of DBMS, OS etc.. All the methods of maintaining the quality of data-repository are composed up of some well defined methods and processes and these steps while deployed over homogenous environment are critically examined in the analysis presented here. The major issues or challenges to be dealt at every stage are highlighted.

Keywords: Data Quality; Data-Repository; Homogenous Environment

I. DATA QUALITY ISSUES

An ultimate objective of having a repository of the data is to provide qualitative data in the hands of its users. Here maintaining, improving the quality of data means modeling and structuring of data in such a way which specify and ensures the freshness, completeness, consistent and noise-free data. The versatile and widely accepted meaning of quality is to have and maintain the data as fit for use and as conformance to requirements [Wang, R. Y. and Strong, D. M. (1996)] [Crosby, P. B. (1984)]. Quality decisions must be based on high quality data otherwise it generates maintenance and repair costs. The cost of deprived data quality must be calculated in terms of user requirements. Even accurate data is of modest value if not interpretable and reachable to the user [Ishikawa, K. (1985)]. Even though, the progress in the context of quality improvement is substantial but the real-world data is usually imperfect. To be interpretable and functional in a proficient and valuable manner, data have to satisfy a set of quality criteria. Data satisfying those quality criteria is said to be of high quality. In general, data quality is defined as an aggregated value over a set of quality criteria [F. Naumann, 2002].

II. HOMOGENOUS ENVIRONMENT

The model of database being homogenous or heterogeneous is determined according to uniformity and compatibility of underlying DBMS. The similarity of operating system (s/w and h/w) is next decisive factor. Within a Homogeneous Environment, data structures of the schema over 'n' sites are at least compatible (if not similar), which also leads in establishing similar query structure for every site. Despite of geographical scattered database the semantics of DBMS is uniform in homogeneous database. Application of various method or techniques at every stage of preprocessing over different sites require the deployment of same mechanisms for cleaning etc. hence leads to similar type of resultant data at each site which actually turns the process of integration, reduction and transformation quite easy to perform. The data cleaning stage gives the data passed through same cleaning procedures at each site and hence same methods would have been used to fill in the missing values, to remove the noise, replacement/insertion of default values etc. Hence, the integration stage get the similar type of data (same default value signifies same data item) from each site for integration making the job quite easier which otherwise would have been different computational techniques for different site consuming more man/machine power. Reduction and transformation becomes less clumsy since the common statistical method may be chosen for each location. Towards the goal of assimilating data quality features into the repository, we illustrate a case in which Table T1 and Table T2 (flat files) have been assumed to be sourced from different but homogenous environment.

Table T1

EMP ID	FNAME	LNAME	SEX	DEPTT	PAY	AGE	D.O.B	PAY_BAND	MARITAL STATUS
RS1001	RAM	SHARMA	M	SALES	20000	24	01.04.1990	7200-400-9600	MARRIED
JK1002	JOLLY	KUMAR	M	MARKETING	25000	203	22.11.1991	7800-500-10300	SINGLE
SA1003	SNEHA	NULL	F	MARKETING	25000	24	15.06.1990	7800-500-10300	SINGLE
KK1004	KARAN	KAPOOR	M	PRODUCTION	18000	26	17.09.1988	5500-300-8500	MARRIED

HS1005	HARSHA	SINGLA	F	FINANCE	28000	27	10.10.1987	8000-400-12000	SEPERATED
MN1006	MOHAN	NULL	M	PRODUCTN	15000	22	06.05.1992	5500-300-8500	SINGLE
NM1007	NEHA	MEHTA	F	SALES	20000	28	18.07.1986	7200-400-9600	DIVORCED
BM1008	BHEEM	NULL	M	MARKETING	26000	25	12.07.1989	7800-500-10300	SEPERATED
KT1009	KIRTI	THAKUR	F	FINANCE	28000	-2	19.12.1987	8000-400-12000	MARRIED
RL1010	RAHUL	NULL	M	SALES	20000	26	28.02.1988	7200-400-9600	MARRIED

Table T2

SSN	NAME	SEX	P.NO.	D.O.B	QUALIFICATION
1001	R.SHARMA	1	+9023-456122	01/04/1990	GRADUATE
1002	J.KUMAR	1	+8763-207630	22/11/1991	SECONDARY
1003	S.KUMARI	0	+7024-619325	NULL	GRADUATE
1004	K.KAPOOR	1	+9182-745014	17/09/1988	MATRIC
1005	H.SINGLA	0	+8132-465726	10/10/1987	POSTGRADUATE
1006	MOHAN	1	+2748-521731	06/05/1992	MATRIC
1007	N.MEHTA	0	+9245-631278	18/07/1988	SECONDARY
1008	BHEEM	1	+8334-598751	12/07/1989	GRADUATE
1009	K.THAKUR	0	+9126-098327	19/12/1987	POSTGRADUATE
1010	RAHUL	1	+8435-619045	28/02/1988	SECONDARY

III. PRE-PROCESSING PROBLEMS OF HOMOGENEOUS DATABASES

Although, the preprocessing part of homogenous databases is quite symmetric in the sense that each stage of preprocessing requires uniform efforts for the cleaning and integrating the data located at each site, yet it is not out of the woods. The preprocessing of homogenous databases is actually surrounded with a small but delicate set of problems. The foremost of them are mentioned here:

A. PROBLEMS FACED WHILE DATA CLEANING

Manually feeding/inputting the data into Database is error prone. Hence, such Databases must be filtered with some proven mechanisms before it is used for inferring any information. [Maletic, J. I., & Marcus, A, 2000] This is also termed as enhancing the quality of data and this aspect is well taken care by Data Cleaning which includes removing the noise from Database & filling in the missing/omitted values.

A.1 Missing values may happen due to non availability of the data-item at the time of entry or availability of illegal/invalid value at the time of entry or some inadvertent change in data afterwards. As evident in table T1, value for LNAME (Last Name) is missing for tuple no. 3,6,8 & 10. Here, missing values are represented with NULL which is not an acceptable standard for a good DBMS. To treat the presence of missing values, if it is opted to ignore the tuple then the information of particular employee is omitted from the database. In case, it is opted to fill this missing value with some arbitrary value may actually escort the noise.

A.2 Noise can occur in our database in different forms like in-consistency or duplicity of data. This noise actually creeps into our database due to insufficient validation procedures or lack of standard check at data entry or due to mishandling of missing data values [Ciszak, L., 2008]. Here, data values for AGE attribute in table T1 for record numbers two and nine are not valid age values (203, -2 respectively). There are a number of methods to choose from, for clearing away the noise. The process of noise removal is time-consuming too.

A.3 To automate the statistically rich methods of data cleaning is in itself a great challenge for every programmer as it requires high programming skills with proper statistical knowledge since the process of filling in the missing values is accomplished with the outcomes (data values), which are usually different when calculated by using different statistical functions. Hence, the use of one method may produce different value as compared to another method. Hence, this again leads to higher level of noise. The success of correctly filling in the missing values largely depends on the abilities of personnel to find the most relevant & most probable data value.

A.4 Expensive resources like memory & CPI: The programmatically or automatic removal of noisy data from Database requires the selection & usage of a statistical method which filters the data effectively [Maletic, J. I., & Marcus, A, 2010]. The chosen method should not only fulfill the purpose of noise removal but should also take the least amount of time since each statistical method for noise removal is very complex in itself & hence requires many expensive resources like memory & CPU time (Cycle Per Instruction), which should be optimized.

A.5 Conflict resolution: The process being manual or automatic, both systems must have to deal with the problem of conflict resolution while filling missing values at different places. The selection of different method

for noise removal may lead to different value to rectify, which needs to be resolved. In many cases, the relevant and available information is actually scarce to remove these conflicts. As shown in following instance (table T1.1 and T2.1), extracted from the given case study, table instances T1.1 and T2.1 do not contain enough information to resolve that whether LNAME of the record may be filled with 'KUMARI' or not since the E_ID are in different formats and the information for field D.O.B is also not available to match the complete record of the employee.

Table T1. 1

EMP ID	FNAME	LNAME	SEX	DEPTT	PAY	AGE	D.O.B	PAY_BAND	MARITAL STATUS
SA1003	SNEHA	NULL	F	MARKETING	25000	24	15.06.1990	7800-500-10300	SINGLE

Table T2. 1

SSN	NAME	SEX	P.NO.	D.O.B	QUALIFICATION
1003	S.KUMARI	0	+7024-619325	NULL	GRADUATE

A.6 Maintenance of cleansed data: Besides above stated issues, maintenance of cleansed data, management of multiple alternate values & cleaning lineage are also constitutes the list of issues of data cleaning. Cleansing lineage is generally kept for the records that have been modified and for those that have been verified within the cleansing process as being correct as well. As soon as, any of the values in the data collection is changed, the cleansing workflow has to be repeated for all those records that include the changed value as part of their cleansing lineage. It is obvious from the above mentioned difficulties of data cleaning that the manual cleaning of database is suitable only for organizations with small databases. For bigger databases or warehouses, the automation of data cleaning is inevitable which in turn has its own problems, visualizing availability of computation resources at minimum cost (in terms of memory & time) still yielding accurate results.

B. PROBLEMS FACED WHILE DATA INTEGRATION

The process of integration of data/instances collected from homogeneous resources involves a sequence of activities starting from acquisition of data across departments/sites (of an organization), finding common keys followed by synchronizing the data in the context of time or alike scales. Although, the process of Integration is well mature yet it is not smooth enough due to below mentioned obstacles. In fact, the each stage of Data Integration is well equipped with its own limitations.[Halevy, A., Rajaraman, A., & Ordille, J., 2006]

B.1 Schema Matching Problems: Schema Matching is a challenging job while integration. Even though the data is collected from homogeneous sources, the type or constraints etc. of data may be heterogeneous [Almarimi, A., & Pokorný, J., 2005].

Finding common key: This may entail to the problem of finding a common key from two or more instances. The problem of finding common key also exists in given case study. The table T1 contains E_ID as its unique field to identify each employee and table T2 contains SSN as its unique key but the data-format and data-type in both of these tables are different hence it can't be deduced with surety that these data-values of such keys refer to same records.

Type heterogeneity: The data types of the fields may also be different including width specifications or pre-defined input pattern etc.[Lim, E. P., Srivastava, J., & Shekhar, S., 1994] e.g. in table T1 & T2 there is a common attribute 'SEX' but one contains value in the form of M/F & other one contains 1/0 for Male/Female respectively. Hence, the problem of data-type heterogeneity arises while integrating data, even in Homogeneous environment. Same problem occurs due to value heterogeneity. As evident from the tables T1 and T2, that T2 contains different value for D.O.B. for record number seven as contained in table T1 for corresponding record; which of these should be assumed to be correct is deceivable.

Semantic heterogeneity: The definition of alike sub-schemas collected from different but similar sources may be different, such as one relation may point to the details of an individual whereas other may specify a family or generic class. In the above case, table T1 has E_ID column as its primary key whereas table T2 has SSN as its unique key to identify its records separately, but actually both of these columns are referring to same employee but with different conventions. Such heterogeneity is difficult to resolve unless the all available information of each record is matched.

B.2 Time Synchronization: This also appears to be a problem (especially in case of temporal Database). Here, lies the problem of finding the compatibility between two different time-periods (such as *workingdays* & *workinghours*). This is necessary to figure out the required level of synchronization between two different but alike fields.

B.3 Socio-logical factors: This plays an important role in successful integration of data since personnel working at different levels have the reluctance to share their data due to various fears such as: loss of power, loss of false fame, non-availability of complete data or meta-data etc.

B.4 Data Integration is AI-complete: The process of Data Integration is actually a good mix of human and machine efforts. In fact, neither only human endeavor is not enough for its implementation, nor only automated solution is likely.[Halevy, A., Rajaraman, A., & Ordille, J., 2006] Even though any such automated application for Data Integration would be only deployer specific. Moreover such system will always give threshold levels of results is also uncertain.

B.5 Miscellaneous technical reasons: Besides all above mentioned problems there are some other miscellaneous hurdles, some of these are only case-specific and rest are general. The job of integration requires proper understanding of sources of data in terms of machine code ASCII or EBCDIC. Semantic meaning of NULL values, need of derived data-fields, need of exact type of join (inner vs. outer). Above all factors the automation of Data Integration has to be concerned for expensive computational resources besides well-defined APIs, flexible list of options to be used with the process of Data Integration

C. PROBLEMS FACED WHILE DATA REDUCTION

Data in Databases or Data Warehouses accumulates at an unprecedented speed. This accumulation further enhances as the size and reach of organization picks up. To effectively utilize the data for mining purposes need arises to for downsizing the data since the accuracy & efficiency of queries degrades rapidly with the size of data. The various techniques of Data Reduction such as aggregation, numerosity reduction or dimensionality reduction are available [Vora, P., & Oza, B., 2013].

C.1 Reduction ability: Data Aggregation contains highlighting & combining various views of data sets in three dimensional representations. And while Attribute Subset Selection (which implements forward selection/backward elimination or their combination), the irrelevant or redundant attributes are removed and hence the size of data is actually reduced. This process of attribute subset selection is required to be applied on table T1, since it contains two attributes AGE & D.O.B., which in fact provides same information i.e age can be calculated from D.O.B. so, reduction is sometimes necessary to improve the quality of data. Dimensionality Reduction focuses on lossless (compression/conversion) reduction of data.[Lei Yu, Jieping Ye, Huan Liu] Here in table T1, the salary is expressed in terms of exact multiples of thousand, so a conversion may be applied over this column where all the values are stored after dividing it with 1000 and hence column name itself may be captioned like SALARY(in thousands). Such a conversion will reduce the size of data being transferred to and fro over disk. In all, there are many methods to choose from. The reduction ability of each method is the important criteria to decide on.

C.2 Loss of interpretability: Dimensionality reduction technique of Data reduction, when applied over datasets may lead to loss of interpretability. For example, in table T1, the attribute PAY_BAND field values represent the actual grade pay which certain employee is drawing and this column values if converted to Pay-band I, II, III or IV etc. as per the grade they fall in, then it will be easier to enter and avoid the possible noise. But this conversion may actually lead to loss of interpretability since it will not be directly meaningful and understandable to each user then. Same happens while interpretation of Principal Components, which may be difficult at times since they do not necessarily correspond to meaningful physical quantities.

C.3 Noise Tolerance: Numerosity Reduction method of Data Reduction, also suffers from the problem of estimating/selecting the random variable which acts as a seed for Regression and log-linear methods of Numerosity reduction. The different selection leads to different level of noise tolerance.

D. PROBLEMS FACED WHILE DATA TRANSFORMATION

Data Transformation is the process of quantitative/mathematical change in variables by applying some very common (variance-stabilizing) statistical functions such as Square Root/Log/Inverse.

D.1 Difficult to validate: Since the mis-formatted data is rampant, the major limitation with the application of Data Transformation using various methods is, one has to be very careful while interpreting the results of transformed data. Extreme, acute, replica or invalid values anomalies arising due to inconsistencies between integrated data sets, constitute the most tedious component of the analytic process, which may actually inflate the cost. Moreover, reformatting and validating data requires alterations that can be difficult to specify and evaluate.

D.2 High skilled labour: The process of transformation requires writing of distinctive scripts in programming languages like Python or R and exhaustive editing is also required using some interactive tool like spreadsheet etc. The usage of these statistical functions requires (automatic/not-by-hand) skilled knowledge of statistics in order to implement these tools in an informed manner.

D.3. Only for anomalous data: one must understand the data prior to performing analysis & must also know the need of exact data analysis technique to be used. Even though there are valid reasons to implement Data

Transformation, still it is not advisable to deploy Data Transformation unless there is actually a valid, clear & legitimate reason [Osborne, J. (2005)]. The impact of particular type of transformation being used such as variations in spelling or formatting, measurement units, or coding schemes (e.g., names vs. abbreviations) should also be understood well. Such as in given tables, the FNAME and LNAME columns of table T1 are to be transformed (actually abbreviated or little variation in presentation) to match the NAME attribute of table T2.

IV. CONCLUSION AND FUTURE DIRECTIONS

All of the data pre-processing techniques which may be applied on homogeneously sourced database are listed in this paper. The application of these scientific and well proved techniques actually raises the quality of data, thus assisting to improve the efficiency and accuracy of text mining, web mining, ubiquitous mining and other data mining activities. Even though these techniques are well established, there are some technological limitations or personnel limitations which urge the usage of these techniques under keen surveillance. The problems which need the attention during implementation of each technique like data-cleaning, data-integration, data-transformation etc. has been presented in this article and these problems also constitute the list of key research issues to be brainstormed on. The inadequate data-quality checks on the process itself, insufficient testing on post-integration datasets, too loose definitions of fields or records, inconsistent interpretation or usage of formats, symbols or codes and finally lack of management attention towards effective pre-processing of data, always keeps the personnel in the vicious cycle of pre-processing.

Although various methods and techniques of data pre-processing have been urbanized, data pre-processing remains an energetic area of research. We believe that outcomes of this research revelation, if properly attended, will give significant strength and ease to the process of pre-processing the homogeneously sourced database.

In this research, we have assumed that the database is piled up from the data collected from similar sources. The next step of our further efforts constitutes the investigation of huge semi-structured or unstructured data creeping in from the web sources which actually escorts the inconsistent or dirty data and also uplifts the complexity of the problem.

V. REFERENCES

- [1] Almarimi, A., & Pokorný, J., 2005, "Schema Management for Data Integration: A Short Survey". Acta Polytechnica, 45(1).
- [2] Ansari, A., Shafie, A. B., & Md Said, A. B., 2012, "Independent Component Analysis Using Random Projection For Data Pre-Processing", International Journal of Computer Science Issues (IJCSI), 9(3).
- [3] Berka, P., & Bruha, I., 1998, "Discretization and grouping: Preprocessing steps for data mining". In Principles of Data Mining and Knowledge Discovery, Springer Berlin Heidelberg, pp. 239-245.
- [4] Ciszak, L., 2008, October, "Application of clustering and association methods in data cleaning". In Computer Science and Information Technology, 2008. IMCSIT 2008. International Multiconference on IEEE, pp. 97-103.
- [5] Gayathri, R., & Malathi, A., 2013, "Exploration of Data Mining Techniques in Record Deduplication". International Journal.
- [6] Halevy, A., Rajaraman, A., & Ordille, J., 2006, September, "Data integration: the teenage years", In Proceedings of the 32nd international conference on Very large databases, VLDB Endowment, pp. 9-16.
- [7] Kaloyanova, K. M., 2005, "Improving data integration for data warehouse: a data mining approach". In Proceedings of the International Workshop: COMPUTER SCIENCE AND EDUCATION, Bulgaria, pp. 39-44.
- [8] Lenzerini, M., 2002, June, "Data integration: A theoretical perspective". In Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, ACM, pp. 233-246.
- [9] Lim, E. P., Srivastava, J., & Shekhar, S., 1994, February, "Resolving attribute incompatibility in database integration: An evidential reasoning approach", In Data Engineering, 1994, Proceedings, 10th International Conference, IEEE, pp. 154-163.
- [10] Lup Low, W., Li Lee, M., & Wang Ling, T., 2001, "A knowledge-based approach for duplicate elimination in data cleaning", Information Systems, 26(8), pp. 585-606.
- [11] Maletic, J. I., & Marcus, A., 2000, October, "Data Cleansing: Beyond Integrity Analysis", In IQ, pp. 200-209.
- [12] Maletic, J. I., & Marcus, A., 2010, "Data cleansing: A prelude to knowledge discovery", In Data Mining and Knowledge Discovery Handbook, Springer US, pp. 19-32.
- [13] Prajapati, D. J., & Prajapati, J. H., "Handling Missing Values: Application to University Data Set", International Journal of emerging trends in engineering and development, 1.
- [14] Teng, C. M., 2001, May, "A Comparison of Noise Handling Techniques", In FLAIRS Conference, pp. 269-273.
- [15] Vora, P., & Oza, B., 2013, "Improved Data Reduction Technique in Data Mining".
- [16] Wang, K., Yu, P. S., & Chakraborty, S., 2004, November, "Bottom-up generalization: A data mining solution to privacy protection". In Data Mining, 2004. ICDM'04, Fourth IEEE International Conference on IEEE, pp. 249-256.
- [17] Witten, I. H., & Frank, E., 2005, "Data Mining: Practical machine learning tools and techniques", Morgan Kaufmann.
- [18] Zhu, B., He, C., & Liatsis, P., 2012, "A robust missing value imputation method for noisy data", Applied Intelligence, 36(1), 61-74.
- [19] Ishikawa, K. (1985). What is Total Quality Control?-the Japanese Way. Englewood Cliffs, NJ. Prentice-Hall
- [20] Wang, R. Y. and Strong, D. M. (1996). "Beyond accuracy: what data quality means to data consumers". *Journal of Management Information Systems*, Vol.12, No.4, pp5-34.
- [21] Crosby, P. B. (1984). Quality Without Tears. New York: McGraw-Hill Book Company.
- [22] F. Naumann, Quality-Driven Query Answering for Integrated Information Systems Lecture Notes in Computer Science, LNCS 2261, Springer, 2002
- [23] Osborne, J. (2005). Notes on the use of data transformations. Practical Assessment, Research and Evaluation, 9, 42-50.
- [24] Dimensionality Reduction for Data Mining-Techniques, Applications and Trends by Lei Yu, Binghamton University Jieping Ye, Huan Liu, Arizona State University.