A Test Data Design Method for Data Quality Evaluation

Zhenyu LIU[1,*], Qiang CHEN[1] and Lizhi CAI[1]

[1]Shanghai Key Laboratory of Computer Software Testing and Evaluating,Shanghai, China

Keywords: Test case, Test data, Data Quality

Abstract. The rapid development of data analyze needs high data quality requirements. The data quality also determine the software quality. This paper proposes a test data deisgn method for evaluating data quality. The paper analyzes the data quality requirement model of typical software applications, considering the typical data quilty problem or issus. According to different evaluation characteristics, the paper proposes a test data method for quality evalaution related to data requirement. The paper also gives detailed for data design to fulfill the evlauation requirements. The method is used to test some data set and found some data issues in actual application databases.

1. Introduction

The software application produced more and more data with the gradually enter the era of big data. The applications start gradually into all areas of society. The data are used to analzye and mine the new information with related technology. Therefore the data have penetrated into all people life from the more and more application, especially the internet appliation. The vast changes for enterprise have become more significant, any hope of success derive value from big data in the enterprise. The society is facing a vast revolution with the data technology and data has become the important basic infrastructure. The features of data in new era consist of the data size, speed, and the dimension of big data. These changes will affect further data processing. The software application must have the ability to handle massive amounts of data which the data size of big data systems. The diversity of software must be able to handle different types of data, including structured, semi-structured and unstructured. The data process is based on the data quaility. The high quliaty will bring the good process results and the poor quality will get the useless outcome. Although process speed, such as data extraction, transformation and loading also important, the data quality should be the first one to be considered. So data cleaning is popluar before the data processing. In this paper, we proposed the data quality evaluaion when data set ready. The data set plays a very important role beforethe data process.

The data set of software application could consists of very large number of structured and unstructured data. The any data analyse process will involve more than one database and will complete in a specific period of time. Due to the low quality and poor system design code, the quality of application as data volume growth will be varied. Even when the amount of data reaches a certain size, the application will crashes and cannot provide proper function service.

The evaluation techniques should meet the quality requirements of data set, espically for fulfill the feature of evaltion model. The quality of dataset will be considered by the data size for data processing diversity. The environment of test data set also cannot meet for software application. Therefore the test project should put forward higher requirements, which requirement not only reflected in many aspects of data size, application processing capability and test environment, but also influence on the test results.

The main structure of this paper is as follows: the next section introduces the model of test case design. The next part gives test data generations. And then illustrated a detailed test project and test results. Finally, there are conclusions and future works.

2. Evaluation Model

The test evaluation model is not considering the influence of test result from data correction factor. However, we study the consistency, conformity and accuracy of data. These quality characters will affect the test result for many dataset in the software application. The undecided data will introduce the some issuses and lead to the unexpected test result. The exceptional data distribution will result in erron results after analyze The typical operation in some big data applcations, such as the input segmentation, redundancy move, sort of operation, should be evlauated before execute the data analysis, such as Map aggregation in Mapreduce.

2.1 Quality Character

However, we not focus on data processing features and also evaluate from a user perspective and inner perspective, such as consistency, conformity and accuracy perspective. In international standard 25012, there are 14 quality characters. These 14 characters consists of accuracy, completeness, credibility, currentness, accessibility, compliance, confidentiality, efficiency, precision, traceability, understandability,availability, portability and recoverability. In this paper, we propose the specified quality characters for typical data set. The basic consistency conformity also be considered the important for the software or information systems.

The consistency focuses on the target entity has values for all expected attributes in software application. The different software or applications provide analysis and processing on the same data set, but the different outcome will be get due to data quality issus. The different of data maybe lead to unexpected analyze results and make the outcome unbelieve.

The completeness focuses on which data associated with a target entity has values for all expected attributes and related target entity instances in a specific context of use. The integrity of dataset is necessary to anlayze. The absense of data maybe lead to analyze fail due to number of data.

The accuracy has attributes that correctly represent the true value of the intended attribute of a concept or event in a specific scenario.

The correctness and completeness in software application are observed for changes from the test result with same data set. The accuacy of test set provides the continuous process ability whether the stable operation with the different software applcation based on same data set.

2.2 Test Case

The test procedure for data quality evaluation is relatively obvious. The main test procedure consists of two steps. The first is loading data set and the second is executing quality evalution , which consists of data process and quality analysis. The data loading is to load test data into data warehouse in specific database or designed test evaluation tools. Data process and analysis is the core operation based on data evaluation algorithm, or rule set.

The test case should be designed from the quality model which described by users or evaluator. The detailed the test procdure and test data should be designed according to the quality character. The different character has the different test method.

The larger data size will increase analyise time, which is not simple linear or exponential. With the data amount of change exponential growth and the data set reaches a certain size, the processing time will be increased under the quality requirements of application.

2.3 Test Data

The test data are the key procedure to produce the different data which not same for source data. The source data will give the actual evaluate result when testers usually observe during the process of quality evalaution. The original data always give the test outcome as same as the test oracle, which developed in the period of test design by senior test developers.

Here we consider the add some special test data in test set to reflect some abnormal data in test data during test design. These abnormal test data will influence the test outcome to some extent. The

abnormal data, such as error data or miss value data will behane influence effect for the end test outcome. The error data will lead to error evlaution result in some circumstance which decided by the software applications. The miss value data will not influence the outcome in normal application. But for lots of software application, the outcome will be different. Therefore, we also need to design some specific data for evalution.

For software application, the variation of data set can impact on the analysis results according to quality of test data. The different test requirement should be considered under the different data set. Here, test strategy is set when the test execution has taken to implement the requirements. The test requirements should convert into quantifiable, measurable, achievable load target to data requirements. The test scenario is selected test data according to different test target. According to data policy, test designers should calculate or designate a variety of direct and indirect target during the test execution.

3. Evaluation Design

The evaluation implement is making the test available after the test design. The available means the test execution will go properly without exception or other unforeseeable consequences. The test purpose of big data software is used to fully analyze the quality characteristics of the application corresponding to test resources.

Test data should be concerned and test cases mentioned in every possible scenario. These conditions must be considered together. In some software, the test data for application features should cover all possible data. Therefore, we need to focus on test data effective. The each typcial test data should be represented the scenario.

3.1 Indicator Design

The indicator design means the number scores for charactics of data quality. Here, we consider the range indicator from 0 to 1. The 0 means the poor data quality, and 1 means the quality meet the requiremnet otherwise. The range from 0 to 1 is not absolute The difference quality requirement is not means 1 represented the high quality. The evaluation indicator after the evaluation will be calculaed. The results should be studied the relation of modify certain parameters becomes large, the distribution changes or trends.

3.2 Data Design

The generation of test data should be representative and inclusive, and in accordance with the degree of data distribution. Data should be representative of the design, including reasonable the value range of test data There are three typical categories for test data, such as noise data, inconsistent data, and duplicate data.

Due to the relatively data quality requirement, the design method requires bonded to each other with a particular quality charactics. For special non-normal data, the data quality need to bypass the data preprocessing. The less strcuture data, the quality should be hard to evaluated. Many applications will be filtered when process with dirty data automatic. The dirty data in many applications not only distorts the data, but also seriously affect the analyze or operational results. In order to make application more accurate, we can design data for specific purpose goal, such as consistent, eliminate duplicate data records. Therefore, some data preprocessing work should be avoided the these abnormal data. But lots of application could be recognized the quality of data and lose of preprocess. The another design method is data modification. The data modification is changing the data in some ways. The ways consists of deletin single record adjust or modifying partly data.

Evaluation design and corresponding to data set will involve the data design and indicator design. The evaluation design is based on the quality requirements and test data, you will not be able to obtain consistency of data management and standardize procedures

4. Evaluation Result

The application or information systems are executed by test case with tested data set. It can be used according to the instructions in demand equivalence classes, boundary value methods. Evaluation may require a lot of data from a centralized select some of them. The basic assumption is that data errors are found. According to the application requirements specification, we use data record row scan method to evalutaion data set. However, existing data or dictionary order method for generate the rules are too simple. Therefore, quaility evaluation is available necessary under certain rules, such as based on regular expression syntax for describe the data. The regular expression which applied data generation can improve the accuracy rate of the data set. The data set can always check directly based on regular expressions. The data set should be targeted structured, semi-structured and unstructured data, particularly those concerned with different data types have certain logic of the data, the data should be carried out according to its logical relationship design.

The three typical result of application-driven instructions are given, for modification from three range from 50% to 80%. However, the work still running, here we only give the few experiments description.

5. Summary

The data quality is an important part of data process application. The result is not only help to get the actual quality, but also help to remove data errors and improve the quality of dataset further. This paper studies influence of test data design method for data quality and introduces the importance of data and consequently carried out evaluation method for data quality. The paper proposes the data design and evaluation technique. Some test results of execution is given finally. The evaluation results of data quality could be used to get the quality score and improve the data quality through the application modification and optimization.

Acknowledgment

The work is supported by Shanghai STCSM Program under Grant No. 16511101202, Shanghai STCSM Program under Grant No. 16DZ1100203 and Shanghai STCSM Program under Grant No. 16DZ1110101.

References

[1] Batini, C., Scannapieco, M. Data Quality: Concepts, Methodologies and Techniques Springer (2006)

[2] Khatri, V. and Brown, C. V. Designing data governance, Communications of the ACM, 53, 1, 148 (2010)

[3] Lizhi Cai. The advent of Big Data: Ready for Software Testing, Software Industry and Engineering, 2013 (5): 15-17(Chinese).

[4] Rabl T, Jacobsen H A. Big Data GenerationM]//Specifying Big Data Benchmarks. Springer Berlin Heidelberg, 2014: 20-27.

[5] ISO/IEC 25012: 2008, Software engineering - Software product Quality Requirements and Evaluation (SQuaRE) - Data quality model

[6] Batini, C., Cappiello, C., Francalanci, C., Maurino, A.: Methodologies for data quality assessment and improvement. ACM Comput. Surv. 41(3) (2009)

[7] Jiang, L., Barone, D., Borgida, A., Mylopoulos, J.: Measuring and Comparing Effectiveness of Data Quality Techniques. CAiSE 2009: 171-185 (2009)

[8] Aiken, P., Allen, M., Parker, B. and Mattia, A.Measuring Data Management Practice Maturity: A Community's Self-Assessment, Computer, 40, 4, 42–50. (2007)

[9] Helfert, M. and F.M.Z. Hossain An Approach to Monitoring Data Quality, in Proceedings of the Americas Conference on Information Systems (AMCIS 2010), August 12- 5, Lima, Peru(2010)

[10] Loshin, D. The practitioner's guide to data quality improvement, Morgan Kaufmann, Burlington, MA(2011)