International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-5, March 2020

Validating a Big Data for Data Quality using Single Column Data Pattern Profiling Technique

K. Makesh Babu, K. Mohan Kumar

Abstract: ***Data quality is important to all private and government organization. Data quality issues can arise in different ways. Due to inconsistent, inaccurate unreliable and loss of data in e-governance, retrieving of accurate data will become a big trouble in decision making. There are some common data quality issues available in a big data. Those issues and causes are cleared by using data profiling. The process of Data profiling methods detects errors, inconsistencies and redundancies in a dataset. Data profiling has different types of analysis techniques to correct the data such as Single Column analysis, Multicolumn analysis, Multi table and Data dependencies. Single column analysis has different set of analysis. In that Pattern matching technique is used to overcome this challenge of inconsistent data along with much needed data quality for analytic results within bounded execution time. Generally pattern matching is performed manually in an organization. Pattern matching helps to discover the various pattern values within the data and validate the values against any organizations. This data pattern profiling method enables to create a valid data set which is used to generate report for future analysis of an organization with more accuracy. This study compares the results of the proposed data pattern logic with other open source tools and proves the efficiency of proposed logic.***

Key words: Big data, Data quality, Data profiling, Pattern matching, Outliers

I. INTRODUCTION

In Big data volume refers to the amount of data generated. It is growing day by day rapidly and it is generated by humans, machines and their interactions on social media. Velocity refers at which different sources generate the data per minute. This flow of data is huge and continuous. Due to the number of users are growing on social media, the data is getting generated daily. The generated data can be in structured, semi-structured or unstructured format.

Revised Manuscript Received on March 30, 2020.
* Correspondence Author
K. Makesh Babu*, Research Scholar, PG and Research Department of Computer Science, Rajah Serfoji Government College, Thanjavur, Affiliated to Bharathidasan University, Trichirappalli, Tamil Nadu, India. E-mail: makeshbabu.k@gmail.com
Dr. K. Mohan Kumar, Head, PG and Research Department of Computer Science, Rajah Serfoji Government College, Thanjavur, Affiliated to Bharathidasan University, Trichirappalli, Tamil Nadu, India. E-mail: tnjmohankumar@gmail.com

Veracity refers to the data is not in correct format or uncertainty of data available due to data inconsistency and in completeness. Data available can sometimes get hard to believe. With many forms of big data, quality and accuracy are difficult to generate qualitative and quantitative measures. Due to uncertainty of data, government or private organization doesn't trust the information they use to make decisions. Poor data quality is possible to decrease the economic status of the organization in any way. So that data quality is must to overcome this issues [1].

Data Quality

It refers to qualitative and quantitative pieces of information. The main characteristics are

- Accuracy: the data generated may be very high or less, its needs to be accurate.
- Relevancy: the data should be really needed to meet the requirements. Irrelevant information causes wrong decisions.
- Completeness: contains enough information in a data set. Missing data causes the statistical analysis.
- Timeliness: the data should be up to date. Out of date information causes loss of time and money.
- Consistency: the data should be reliable. It ensures the data collected in an expected format.
- Granularity: the data should be sufficient and detailed.
- Uniqueness: the data should be distinct. Redundancy causes wastage of memory space.[2].

Classification of Data Quality Issues

There are four different data quality issues

1. Data Quality Issues at Data Sources
2. Data Quality Issues during the Extract, Transform, Load
3. Data Quality Issues during Data Modeling
4. Data Quality Issues in Schema Design

In the above data quality issues the first issue is most important.

Data Quality Issues at Data Sources

Large data generates from different sources like interpersonal data entry, manual files, or through another database. Once the data is loaded into the new database, it needs to identify and fix such issues. In other way various files are combined to generate a single file, resulting in data quality issues. As the part the possible causes of data quality issues occur only at the source stage [1].

Common Data Quality Issues

Some general ways of issues occur in data quality which are given below.

1. Poor Organization

If data is not able to easily search, then it becomes significantly more difficult to make use of it. Different organizational methods and procedures are used to represent the data in dozens of ways.

2. Too Much Data

Government and other organization related to the public can generate more volumes of relevant information and it is also important to their daily needs. So too much data is generated and stored inside the database. While it might seem like "too much data" can never be a bad thing, more often useful and useless data may stored.

3. Manual Data Entry Errors

Data entered by the humans may have errors without knowing to them. Data entered in the wrong field, missed entries and so on are virtually inevitable. This is the very first sources of errors.

4. OCR Errors

Machines can make mistakes when entering data. Most of the organizations must digitize large amounts of data quickly, so they use Optical Character Recognition technology to do so. OCR technology is used to scan images and extracts the text from them. The problem with OCR is that the data entered is always not perfect.

5. Inconsistent Data

When dealing with multiple data sources, inconsistency is a big data quality problem. The repeated records might present number of times in a database. Redundancy data is one of the biggest problems that reduces the consistency and wastage of time and space than any other data issue.

6. Poor Data Security

Security is also one of the biggest problems in an organization. It is failed to handle their data in a professional and secure manner. People data must always be precautions in place to make sure that it can't be used for theft, fraud and spam.

7. Poorly Defined Data

Data is defined not correctly then it causes complexity in analysis for making decision and creates problem in the management. Extracting the data according to the needs is also difficult.

8. Incorrect Data

Data entered incorrectly or not maintained correctly or entered in the wrong field that creates an issue. If the data is not normalized as per the system of records then there will be a biggest issue that occurred in an organization.

9. Lack of Complete Information

When compiling a data set error may occur frequently because of not having all information. That is missing of information in databases.

10. Ambiguous Data

Data entered in a database may be wrong. In a phone id column if the data entry may be longer than the typical ten digits or some of the alpha numeric data or some other data is entered then it is hard to answer quickly with a large body of data.

11. Duplicate Data

Two or more data entries are completely identical then it causes redundancy. If the two telephone numbers are same in a single column that creates an issue. So sorting out duplicate entries makes the best use of data for processing.

12. Data Transformation Errors

While converting data from one format to another error can happen. Interpreting of data is must because data structure from one database is different from other. Keeping the data in a common format is easy to analysis [2].

Above issues should be corrected and stored for reporting and decision making. This is done by only data profiling.

Data Profiling and their Techniques

Data profiling analyze and asses the big data to discover, understand and collecting statistics information about data. This knowledge is then used to improve data quality as an important part of monitoring and improving the accuracy.

The four general methods of data profiling techniques which give better data quality are: Single column profiling, Multi column or Cross-column profiling, Multi table or Cross-table profiling and Data rule validation [3].

Column Profiling method has three stages, which are Analysis of column based on according to properties, according to its statistics functions and according to dictionary. This method can be useful to find frequency distribution and data value patterns within a column of data to find the maximum errors.

Cross-column Profiling discovers the column and column combination with unique values that are defined. It is made up of two processes: Key analysis and Dependency analysis. Key analysis examines collections of attribute values with the possible primary key. Dependency analysis is a difficult process that determines whether all columns are related to that dataset. Both techniques help to analyze dependencies among data attributes within the same table.

Cross-table Profiling uses foreign key analysis, which compares all columns in all selected tables. It identifies the different columns that share the common domain. It indicates the relationship between the common domains.

At last, **Data rule validation** uses data profiling to verify the data instances and data sets that are valid to predefined rules. This process helps to find the ways to improve data quality[4]. The following Figure 1 shows the complete sketch of data profiling.

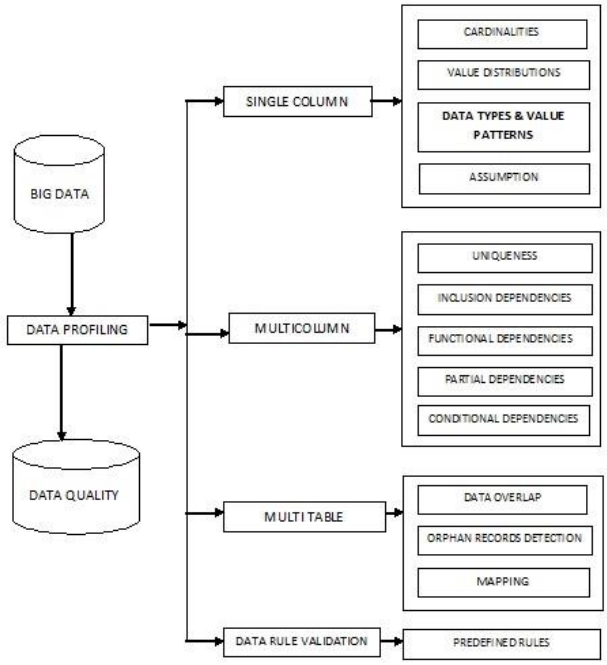

Fig. 1.Types of Data Profiling Techniques

Single Column Analysis

Single column profiling refers to analysis a value in a column and values stored in a single column are independent of all other columns. In the Fig.1 it shows the types of data profiling techniques and the data pattern method is highlighted in the single column techniques. This technique scans through a table and counts the number of times each value shows up within each column. This method also can be useful to find data value patterns within a column of data. Column profiling provides statistical measurements within a single column. Some of the single column analysis performed using Cardinalities, Values Distribution, Data types and Value patterns [5]. This study mainly deals with data patterns.

Data Patterns

Data value pattern is an important aspect in Single column analysis, so this is to be corrected first. Data patterns determine if the data values in a single column are in expected format. This technique quickly validate whether the data is consistent across the data source. An outlier is a pattern, value or frequency for a column excluded from the dataset. This method detects the outliers in the summary view of profile results to identify different patterns, values, and frequency that fall outside the expected range of values.

Outliers and Redundancy is often caused by human error, such as errors in data collection, recording, or entry [4]. The main objective of this research work is to identify the data patterns and analyze the benefits.

II. LITERATURE SURVEY

Suraj Juddoo [1] the paper referred a method for building data quality rules, and how data cleaning potentially involved in data transformation. Data quality is a major important for decision making in all aspects. Also this paper proposed a data profiling tool frame work important for big data to data quality.

S R Amethyst et al., [2] focused on Alphanumeric Pattern profiling algorithm. Data Integration proves that Alphanumeric Pattern profiling algorithm is one important factor for data profiling, which data profiling can find data that are not fit into business rules. It enables companies to clean their data according to the business rules they implementing on their business process.

Brett Dour and Pat Herbert, [5] pointed out the five building blocks of data management they are data profiling, data quality, data integration, data augmentation and data monitoring. Also it presents the structure discovery analysis in that how data pattern format specified using Data profiling techniques.

Tien Fabrianti Kusumasari and Fitria, [6] describes techniques of profiling analysis using open source tool openrefine. Also deals with the data techniques used are binning, smoothing, filtering and data clustering.

III. METHODOLOY

In an organization bad data comes from data receiving from various sources of input. It could be the data sent from another organization, or in many cases, data collected by different software. Therefore, its data quality cannot be guaranteed. A good data profiling tool has a capable of examining the aspects of the data format and data patterns

It is also essential to automate the data profiling and data quality alerts so that the quality of incoming data is consistently controlled and managed [5].

Data Pattern in Data Profiling

Analysis of Data Pattern in a column is done using Data Pattern Profiling algorithm from the previous research performed by some application, which gives less pattern matching and percentage of error recovery is less.

The proposed Data Pattern profiling algorithm finds more pattern matching and user can correct error easily. The computation required for finding and replacing character into a pattern is referred as Replace String. The data collected from an organization is imported to a file with extension of .xls, (Excel file) is used as input [6],[7].

Data Pattern Profiling Algorithm

The following steps explain the flow of the Data Pattern Profiling algorithm to correct phone numbers stored in a column.

Step 1: Transform all numeric text to ‘9’,
Step 2: Transform all uppercase alphabets to ‘U’
Step 3: Transport all lowercase alphabet text to ‘L’
Step 4: Transport all special character to ‘X’
Step 5: Trimming the column data (i.e., remove all blank space in the text and replace by B)
Step 6: Finally Sort the records and group by all records.

Replacing a string method replaces the character or sets of characters in a column with another character, and regular expression can be used in this method.

At last group by (i.e. filters) method is used to group all the patterns. So this method identifies how many patterns are generated. It also finds redundancy and empty value in a single column may cost loss of any telephone numbers in a column. Equal string method finds the same data occur in a column and it is referred to as redundancy. When the particular column defined as null then there is a possibility of leaving the field as empty at data entry level. So there is a chance of empty value or missing value or no value occurs in a single column.

Evaluation and Discussion

The Dataset used in this research is from private organization dataset. This organization regularly sends short message services to the registered phone numbers. There is a column named 'teleid' which contain telephone numbers of the users. The data type and size of this id is alphanumeric and 10 character set. Nearly 2000 records of data were imported to open platform and by copy-pasting data into excel in which later processed with data pattern profiling algorithm.

In previous research data cleaner identifies the pattern for clearing the errors. The main constraints are column should not be empty and must be unique. Previous methods identify only fewer patterns. The proposed method gives more patterns than the previous methods.

Table-I: Results of Outliers using Data Pattern Profiling

Patterns	Total records	Remarks
9999999999	1890	Correct values
9B99999999	29	Blank space
99L9999999	15	Small letter
99U9999999	19	Capital letter
999X999999	11	Special character

Table-I Shows the different set of patterns generated by this proposed data pattern profiling algorithm. Also it counts the total group of records of each pattern. This table shows only 1890 records are having a correct telephone numbers and the remaining patterns are defined to be errors.

Table-II: Results of redundancy

Column name	Total Redundancy	Empty value
Tele id	23	3

The above Table-II shows how much redundancy occurred in this column. According to the table II 23 phone numbers are entered wrongly. In this case the information send to the redundancy numbers creates wastage of expenses. The data entry missing in a phone id column for the users referred as empty value. In this case user cannot receive any messages. It makes loss of information to those 3 users.

Results comparisons with Data Cleaner

The following Table-III shows the comparison result of this proposed method and the previous method.

Table-III: Patterns Metrics

Methods	No. of different Patterns	Total Patterns
Data cleaner	4	45
Data pattern Profiling	5	74

The implementation is done by using dataset contained in Excel files while the output result can be saved into a database for further action. The Data Cleaner method only serves a preset function that cannot be customized according to our needs. Table-III shows that how data cleaner cannot recognize some error patterns occurred in these formats. This data pattern profiling algorithm finds more patterns in a column.

Fig. 2. Shows the percentage of patterns generated using data cleaner and data pattern profiling process is executed using some sample data. It clearly proves that the proposed algorithm identifies more patterns and thus provides the opportunity to clear more errors compare with the previous method.

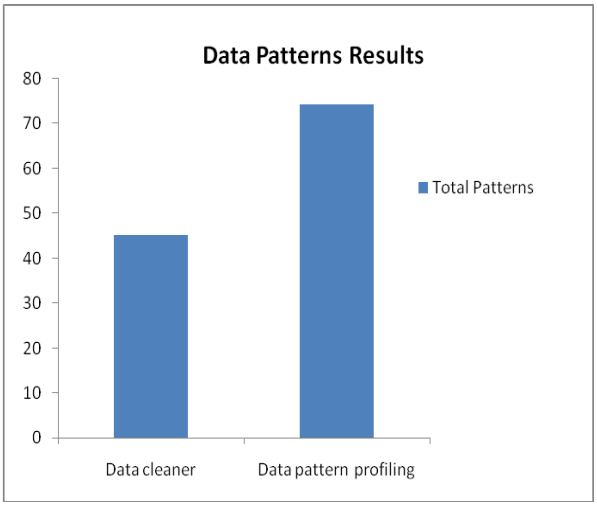

Fig. 2.Comparison of algorithms

Benefits of new Method in Data Profiling

1. Sending short message services to an invalid telephone number is reduced.
2. Sending short message services to a same number is also reduced due to finding redundancy in a column. The impact of this is loss of cost.
3. Identifies the phone number which short message service is not send. Due to the loss of communication the organization income may be loss or delayed.

IV. CONCLUSION

Data profiling proves to be an important step for data quality management and it ensure that the data stored is fit for the data governance. The maximum causes occurred at data sources only. Data profiling is much needed step to get good quality of data. The transformation of data pattern profiling algorithm and the implementation gives more patterns than the other open source tools. An occurrence of more patterns in a column indicates the errors so the user can easily correct the errors manually.

Retrieval Number: E2749039520/2020©BEIESP
DOI: 10.35940/ijitee.E2749.039520
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

The enhanced data profiling method is most important to improve the data quality. It should find more errors and also it reduce time and cost. So, this methodology will give better performance in data profiling technique.

REFERENCES

1. Suraj Juddoo, "Overview o data quality challenges in the context of Big Data," International Conference on Computing Communication and Security (ICCC), 2015, DOI: 10.1109/CCCS.2015.7374131.
2. Joachim Schmid, "The main steps to data quality," Springer 2004, ICDM 2004, Advances in Data Mining, PP. 69-77
3. Felix Naumann-quatar, "Data Profiling Revisited," SIGMOD Record, December 2013 (Vol. 42, No. 4), DOI: 10.1145/2590989.2590995, PP. 40-49.
4. Ana Rodrigues, Av.Pro. Cavaco Silva, "Data Profiling : Identification of Data problem through data analysis," Instituto Superior Technico-Campus Tauspark, 2970-880 Porto Salvo, Portugal.
5. Brett Dorr, Pat Herbert, "Data Profiling: Designing the blueprint for improved data quality," DataFlux corporation, Cary, NC. paper 102-30, 2005
6. Tien Fabrianti Kusumasri, Fitria, "Data Profiling for Data Quality improvement with Openrefine," ICITSI, Bandung – Bali, October 24 – 27, 2016, ISBN: 978-1-5090-2449-0.
7. S R Amethyst1, T F Kusumasari1 and M A Hasibuan, "Data pattern Single column analysis for data profiling using an Open Source Platform," 453, ICDECS, 2018

AUTHORS PROFILE

K. Makesh Babu pursed Master of Computer Application from Bharathidasan University, Thiruchirappailli and M.Phil from Manonmaniam Sundaranar University, Thirunelveli. Currently doing Ph.D as a part time research scholar in PG and Research Department of Computer Science, Rajah Serfoji Government College, Thanjavur, Affiliated to Bharathidasan University, T.N, India. He is having 18 years of teaching experience. His main research work focuses on Big Data Profiling.

Dr. K. Mohan Kumar received Master of Computer Science, Ph.D in Computer Science from Bharathidasan University, Tiruchirappalli, and M.Phil computer science from Manonmaniam Sundaranar University, Thirunelveli, India, currently working as Head in PG and Research Department of Computer Science, Rajah Serfoji Government College, Thanjavur, T.N, India. His main research work focuses on Network Security, Machine Learning and IoT, published more than 50 research papers in reputed International journals. He has 25 years of teaching experience and 20 years of Research Experience.