IOP Conference Series: Materials Science and Engineering

PAPER • **OPEN ACCESS**

Data Pattern Single Column Analysis for Data Profiling using an Open Source Platform

To cite this article: S R Amethyst *et al* 2018 *IOP Conf. Ser.: Mater. Sci. Eng.* **453** 012024

Data Pattern Single Column Analysis for Data Profiling using an Open Source Platform

S R Amethyst[1], T F Kusumasari[1] and M A Hasibuan[1]

[1] Information System Department School of Industrial Engineering, Telkom University Bandung, Indonesia

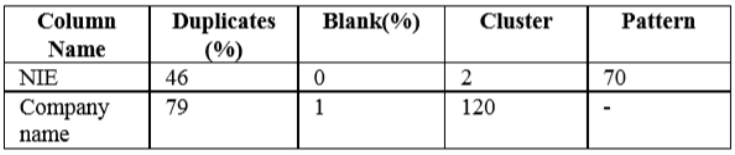
Abstract. The importance of data quality might have a major impact on the company's existing business processes. But there are still many companies that yet to understand the importance of data quality. Many cases that often occurs to the quality of data in many companies in Indonesia is that the inputted data are not filtered, so there are issues about not standardized data pattern. This case can be handled with data preprocess in which one of the methods are data profiling. Data profiling is a proses of collecting an information of a data. In this research the main focus of the analysis by conductin data profiling using data pattern method and algorithm that adopting from OpenRefine and then modified. The results of the profiling using open source tools Pentaho Data Integration, Google OpenRefine and Data Cleaner are really difference, while Pentaho Data Integration and Google OpenRefine found exactly 70 data patterns, Data Cleaner only find 31 data patterns.

1. Introduction

Data is a raw fact that are yet to be processed, which are not directly can be understand by said data receiver. So the data needs to be processed first before finally serve as an information to be receive and understand and used by company for supporting its business process [1]. Globally the sum of inaccurate data has rising from 17 percent to 22 percent just in time of 12 months, and companies in United States believe that they have the highest inaccuracies percentage which is 25 percent. The main reason of data inaccuracy is human error. The level of data inacuraccies are related to lack of sophisticated data quality management strategies [2]. A research by Bachard Pace in 2011 shows that from 195 undergraduated students average number of errors by students that inputted single entry without checking it is as high as 12.03 while the students who did double entry receive average number of errors as low as 0.34 [3].

Previous research performed by Tien and Fitria using open source tool Google OpenRefine in one case in BPOM where the data being process are NIE (Number of Edible Permits) and Company Name. In this research the business rule applied are that NIE cannot be empty, must be unique and have similarities in alphanumeric patterns. The result of the research shows that the NIE has 70 patterns on 5000 rows of data. Duplication analysis needs to be combined with other elements because one production with a single license number can be duplicated if the factory location, volume and weight of a package are different [4].

Data pre-process is every action taken before the data analysis process begins, in which data analysis can give a better insight about the studied subject. The purpose of the data analysis is to gain knowledge that can be used as a decision making guidance. To be able to find the problem to the data, a prevention must be performed by using data analysis tools to produce acceptable data [5].



Column Name	Duplicates (%)	Blank(%)	Cluster	Pattern
NIE	46	0	2	70
Company name	79	1	120	-

Figure 1. Data profiling result using Open Refine

Data profiling are a set of activities and processes to find a metadata from a given dataset, with the purpose of ensuring data accuracy and consistency, and finding data duplication therefore the real value of a data can be received and can be used for decision making [4]. Data profiling perform to a one column is called single column profiling. Single column profiling refers to analysis a value in a column, and has ranges from simple calculations and aggregation functions to distribution analysis and the discovery of patterns and data types [6].

Data profiling are an important aspect to reach Data Quality Management, Data Warehousing and Master Data Management, in which all of them are part of the Data Governance. Data governance is planning, oversight, and control over management of data and the use of data and data-related resources [7]. Data governance involves processes and controls to ensure that the information from each gathered and inputted character by organizations are true, accurate and unique [8].

In this research, many cases that often occurs to the quality of data in many companies in Indonesia is that the inputted data are not filtered, so there are issues about not standardized data pattern. Thus affecting the quality of a data. A bad data quality can affect data governance. It is necessary that the data is clean in fulfilment of master data management strategy, in which a clean, unique and standardized data is required to perform data warehouse. This study uses an open source tool that refers to Google OpenRefine with logic implemented will be compared with other open source tools.

2. Data Pattern Algorithm

A common and useful result from data profiling is frequent patterns extraction of a data from a column. Data that are not fit to such a pattern are likely ill-formed or not standardized. A challenge when determining patterns is to find a balance between generality and specifity [9]. Data pattern will be used as decision making guidance [4]. There are few ways to extracting a data pattern few of them are Minimal Description Length for Potter wheel tool, The ReLIE search algorithm and an Alphanumeric pattern profiling.

Minimal Description Length principle is a method for inductive inference that provides a generic solution to a model selection problem. The insight that MDL based on is that “any regularity in the data can be used to compress the data, that is to describe it using fewer symbols than the number of symbols needed to describe the data” [10].

ReLIE is very powerful regex learning algorithm and very effective for certain classes of entity extraction, particularly under conditions of cross-domain and limited training data [11]. ReLIE search algorithm is designed for information extraction from textual data. The algorithm creates regular expressions based on training data with positive and negative examples by systematically transforming regular expressions [9].

Alphanumeric Pattern Profiling is used to determine alphanumeric pattern of a certain column and the row count for each pattern using Regular Expressions. Regular Expressions (regex) is a pattern that regular expression egine attempts to match in inputted text. A pattern consists of one or more character literals operatior or constructs [4].

Based on the Minimal Description Length, ReLIE and Alphanumeric Pattern Profiling, the researcher choose the Alphanumeric Pattern Profiling due to the similarity of the flow and how it can be implemented in this case using Pentaho Data Integration.

The illustration of process flow of the Alphanumeric Pattern Profiling algorithm will be mapped on Figure 2.

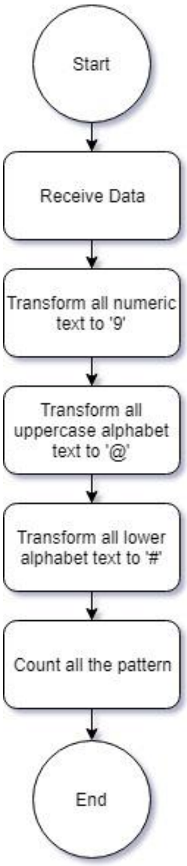
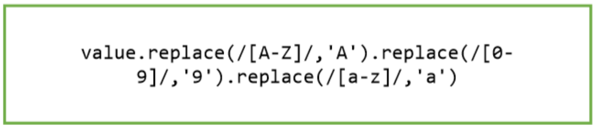

Start
Receive Data
Transform all numeric text to '9'
Transform all uppercase alphabet text to '@'
Transform all lower alphabet text to '#'
Count all the pattern
End

Figure 2. Alphanumeric Profiling Pattern algorithm flow illustration

Figure 2 shows the algorithm pattern that going to be implemented in Pentaho Data Integration. The algorithm starts with getting the data that going to be processed. After the data received, then we choose the column in which the data will going to be processed. Next, the algorithm will transform all numeric text in the column to a specified pattern, in this case the character is ‘9’. After the numeric, now all of the uppercase alphabet will be transform into specified pattern, in this case the character is ‘@’. The last transformation would be the lowercase alphabet, it would be transformed into specified pattern, in this case the character is ‘#’. The last step would be counting all the pattern that has produced by previous steps.



```
value.replace(/[A-Z]/,'A').replace(/[0-
9]/,'9').replace(/[a-z]/,'a')
```

Figure 3. Regex used by Tien and Fitria for Alphanumeric Pattern Profiling [4]

Alphanumeric Pattern Profiling algorithm was used in previous research performed by Tien and Fitria which was performed using the following regular expression (regex) as seen in figure 3. This regular expression works by replacing the character found in a column into a certain character. The first value.replace(/[A-Z]/, ‘A’), is replacing all of the uppercase alphabet character of a data in a column into character ‘A’. The next is .replace(/[0-9]/, ’9’), is replacing all of the numeric character of a data in a column into character ‘9’. The last is .replace(/[a-z]/, ‘a’) is replacing all of the lowercase alphabet character of a data in a column into character ‘a’.

3. Method

The research method used to find data pattern in a column is divided into 3 stages, same as research method used by Febri in the previous research [12]. The first stage of this research method is mapping the function of the Alphanumeric Pattern Profiling algorithm logic to Pentaho Data Integration function. The second stage is design and configure the functions used in Pentaho Data Integration. The last stage is Evaluation, analysis and comparison between Pentaho Data Integration results with Data Cleaner and Google OpenRefine.

The first stage of the research method is perform mapping the algorithm into Pentaho Data Integration Functions by analyzing the flow of the algorithm which can be seen in Fig. 2 and customize its component on Pentaho Data Integration. Data Pattern method focus on finding the pattern of a alphanumeric data.

The second stage is designing and configuring the algorithm into Pentaho Data Integration component. Transformation configuration is performed gradually for each step according to the Alphanumeric Pattern Profiling algorithm and tested in every step made. After the test result of every step is accordance with the existing transformation then proceed to the third and final stage.

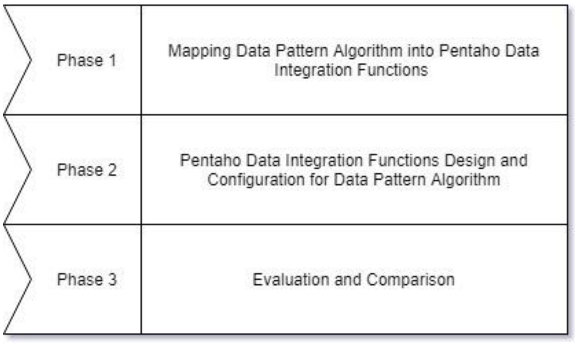

Phase 1
Mapping Data Pattern Algorithm into Pentaho Data Integration Functions
Phase 2
Pentaho Data Integration Functions Design and Configuration for Data Pattern Algorithm
Phase 3
Evaluation and Comparison

Figure 4. Flow of implementation using Pentaho Data Integration

The final stage is performing evaluation and comparison of the result of the Pentaho Data Integration transformation with the result of Data Cleaner and Google OpenRefine. The evaluation is done with the same amount of data and the same flow process. For this research the data used are from Indonesian Government Agencies and are Excel file type.

4. Algorithm for Data Pattern in Data Profiling

Analysis of Data Pattern in a column is done using Alphanumeric Pattern Profiling algorithm from the previous research performed by Tien and Fitria which can be done with various application. In Pentaho Data Integration the component used for finding and replacing into a pattern is called Replace String.

The flow of the Alphanumeric Pattern Profiling algorithm can be seen in figure 2, which will be mapped according to the flow in Pentaho Data Integration tool described in table 1. For Receiving Data, the component used in Pentaho Data Integration is Excel File Input, which can be used to import data from any files with extension .xls. For transform all numeric text to ‘9’, transform all uppercase alphabet text to ‘@’ and transform all lowercase alphabet text to ‘#’, the component used in Pentaho Data Integration is Replace in String. Replace in String can be used to find replace any character or sets of a characters in a column with another character, and regex can be used in this component. The last is count all the pattern, the component used in Pentaho Data Integration is Group By, which can be use calculate values over a defined group of fields.

Table 1. Alphanumeric pattern profiling algorithm mapping to Pentaho Data Integration components

Algorithm	Pentaho Data Integration Component
Receive Data	Excel File Input
Transform all numeric text to '9'	Replace in String
Transform all uppercase alphabet text to '@'	
Transform all lowercase alphabet text to '#'	
Count all the pattern	Group By

For Data Pattern logic implementation into Pentaho and the component used can be seen in Figure 4.

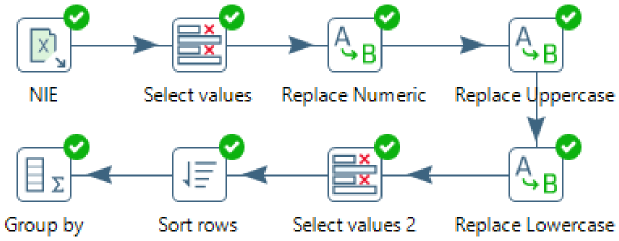


Figure 5. Logic Implementation of Data Pattern Algorithm.

The configuration for the components can be seen in table 2. The Excel File Input is used to get the data from an .xls files. And then the data will be passed to Select Values to determine which column will be processed for pattern profiling. In this component will select NIE for processing. Then the NIE column will be passed into multiple Replace in String components, the first is Replace Numeric, then Replace Uppercase and then Replace Lowercase. For the Replace Numeric, the component will find any numeric character through a regex '/d'. This will find all numeric character in a given column, and then replace it with character '9' with the configuration Replace With, and then save it in a new column named PATTERN Numeric. For the Replace Uppercase, the component will find any uppercase character through a regex '[A-Z]' in PATTERN numeric column. This regex will find all uppercase alphabet character in a given column. then replace it with character '@' through Replace With configuration and then save it in a new column named PATTERN Uppercase. And the last is Replace Lowercase, the component will find any lowercase character through a regex '[a-z]' in PATTERN Uppercase column. This regex will find all lowercase alphabet character in the given column. then replace it with character '#' through Replace With configuration and then save it in a new column named ACTUAL PATTERN. After all Replace in String components finished their works, the ACTUAL PATTERN will be send to Sort By component to be sorted ascendingly. This step is required for the Group By component required a sorted column to be able to process it. The last component used is

Group By. This component will count every pattern and then group it based on the pattern itself, thus the sum of the pattern can be found.

Table 2. Configuration and description of PDI Components

PDI Component	Functions	Configuration
Excel File Input	Importing data from .xls files	-
Select Values	Selecting column that will be found the pattern	Fields : NIE
Replace Numeric	Transform all numeric text to '9'	Regex : \d In Stream : NIE Replace with : 9 Out Stream : PATTERN Numeric
Replace Uppercase	Transform all uppercase alphabet text to '@'	Regex : [A-Z] In Stream : PATTERN Numeric Replace with : @ Out Stream : PATTERN Uppercase Case Sensitive : Y
Replace Lowercase	Transform all lowercase alphabet text to '#'	Regex : [a-z] In Stream : PATTERN Uppercase Replace with : # Out Stream : ACTUAL PATTERN Case Sensitive : Y
Sort By	Sorting the data based on the pattern	Ascending
Group By	Counting all of the pattern	Aggregates : ACTUAL PATTERN Type : Number of Rows

The implementation is done by using dataset contained in Excel files while the output result can be saved into a database for further action.

5. Evaluation and discussion

The Dataset used in this research is from Indonesia Government agencies dataset and the column used is NIE (Nomor Izin Edar) which the researcher obtained from the url http://cekbpom.pom.go.id. Then 5000 records of data were imported to OpenRefine by copy-pasting data into OpenRefine import page, and then exported into an .xls files in which later processed with Pentaho Data Integration and Data Cleaner. NIE is a number that identifies the permit to distribute food across the country. The business rule applied in the government agency is that the NIE column must not be empty, must be unique to each of their entity, and having similarity in alphanumeric pattern. If the NIE column was empty, it means that a product doesn't have a permit therefor violating the laws of traditional drugs distribution in the country. Second, NIE must not be the same or must be different for each entity, because one NIE must shows one certain product in a certain packaging. And the last, NIE have standard pattern rues that are established in a certain time period, showing standardization that applies nationally for multi-region agency.

The Dataset used in this research is from Indonesia Government agencies dataset and the column used is NIE (Nomor Izin Edar) which the researcher obtained from the url http://cekbpom.pom.go.id. Then 5000 records of data were imported to OpenRefine by copy-pasting data into OpenRefine import page, and then exported into an .xls files in which later processed with Pentaho Data Integration and Data Cleaner. NIE is a number that identifies the permit to distribute food across the country. The business rule applied in the government agency is that the NIE column must not be empty, must be unique to each of their entity, and having similarity in alphanumeric pattern. If the NIE column was empty, it means that a product doesn't have a permit therefor violating the laws of traditional drugs

distribution in the country. Second, NIE must not be the same or must be different for each entity, because one NIE must shows one certain product in a certain packaging. And the last, NIE have standard pattern rues that are established in a certain time period, showing standardization that applies nationally for multi-region agency.

The result of the test comparison between open source tools Pentaho Data Integration, Data Cleaner and Google Open Refinehave a far difference in Data Pattern method, as seen in figure 5. The difference are caused by Pentaho Data Integration can implement the algorithm that was used in Google OpenRefine by previous research by Tien and Fitria which is proved to be very detailed in finding data pattern. The algorithm used in Pentaho Data Integration and Google OpenRefine is turning every character in a data into a specified pattern while Data Cleaner also do the same using the Pattern Finder function but Data Cleaner will group a data that have almost similar pattern into one category. Pattern Finder function in Data Cleaner are not fully customizable, therefore we cannot implement the Alphanumeric Pattern Profiling used in Pentaho Data Integration and Google OpenRefine in Data Cleaner. Due to that reason, Pentaho Data Integration and Google OpenRefine can find more pattern than Data Cleaner did.

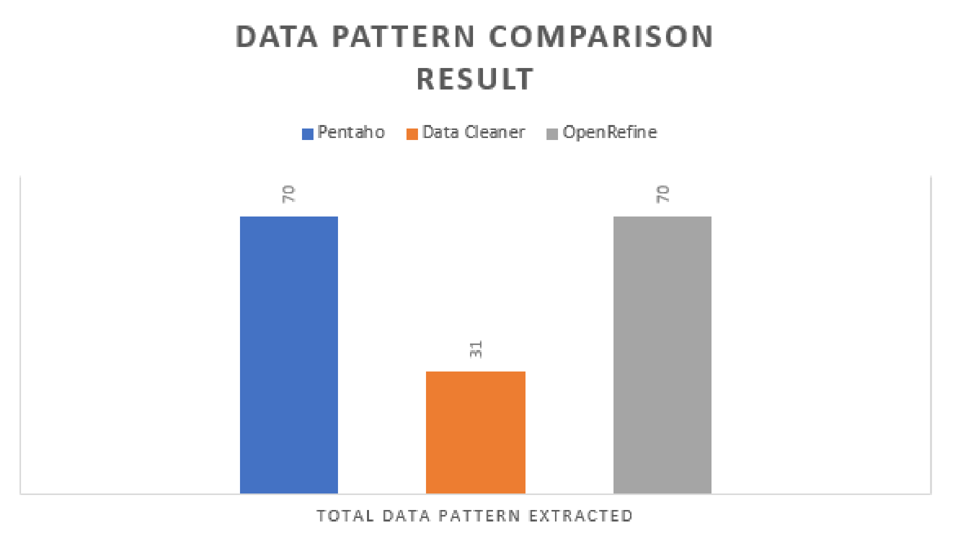


Figure 6. Result Comparison for Data Pattern Method

The final result of the comparison is that Pentaho Data Integration found 70 different patterns the exact same as final result that found by OpenRefine while Data Cleaner only found 31 different patterns. The most noticeable result difference is that there are some data pattern that cannot be extracted by Data Cleaner and it only shown as question mark (?), and this thing doesn't occur in Data Pentaho Integration using Alphanumeric Pattern Profiling algorithm. Other difference is that Data Cleaner tend to group data pattern that has a longer length that have the same pattern with data with shorter length into one group, this resulting in not detailed pattern profiling. This thing cannot occur in the Pentaho Data Integration and Google OpenRefine because the Alphanumeric Pattern Profiling algorithm, convert every character into a specified pattern one by one, so the result can be more detailed and more vary. This result points that Pentaho Data Integration can be very detailed according for the needs for data pattern profiling and it also points Pentaho Data Integration can find that the problems found in the tested dataset is an companies problem where a lot of data not standardized accordingly to the business rule.

The differences found in the result of data pattern extraction in Pentaho Data Integration and Data Cleaner because of the Alphanumeric Pattern Profiling algorithm that used by previous research by Tien using Google OpenRefine, can be implemented in the Pentaho Data Integration. The detailed flow that can be applied to Pentaho Data Integration component according to the algorithm. The same thing cannot be done to Data Cleaner because Data Cleaner only serve a preset function that cannot be customized according to our needs.

6. Conclusion

Transformation by performing data pre-process with data profiling process and the implementation of Data Pattern method using Alphanumeric Pattern Profiling algorithm and implemented using Pentaho 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. enable companies to clean their data according to the business rules they implementing on their business process.

Data profiling proves to be an important aspect for data quality management, to ensure that the data quality is fit for the data governance. Data profiling also profile any data that are not fit into their business rules, if they are not fit into their business rules, then it is necessary to perform data cleansing which occur before building Master Data Management and Data Warehousing which is part of the data governance. Data profiling ensuring data quality to fit for these things. If the data are clean then, Master data management and data warehousing will be possible and therefore Data Governance is fulfilled.

Recommendation is to continue this data profiling to data cleansing and data monitoring. So master data management and data warehousing can be ensured therefore data governance are possible.

Acknowledgments

This research is supported and funded by PPM Telkom University on behalf of Information System research group in data science.

References

[1] M. Lebied, “The 5 Pillars of Effective Data Quality Management (DQM).” 2017.
[2] B. Desai, “The state of data,” *Proc. 18th Int. Database …*, pp. 77–86, 2014.
[3] K. A. Barchard and L. A. Pace, “Preventing human error: The impact of data entry methods on data accuracy and statistical results,” *Comput. Human Behav.*, vol. 27, no. 5, pp. 1834–1839, 2011.
[4] T. F. Kusumasari, “Data Profiling for Data Quality Improvement with Openrefine,” 2016.
[5] A. FAMILI, W. SHEN, R. WEBER, and E. SIMOUDIS, “Data preprocessing and intelligent data analysis,” *Intell. Data Anal.*, vol. 1, no. 1–4, pp. 3–23, 1997.
[6] Z. Abedjan and F. Naumann, “CONFERENCE: Data Profiling,” pp. 1432–1435, 2016.
[7] P. Cupoli, S. Earley, D. Henderson, and Deborah Henderson, “DAMA-DMBOK2 Framework,” p. 26, 2014.
[8] A. Yulfitri, “Modeling Operational Model of Data Governance in Government,” 2016.
[9] Z. Abedjan, L. Golab, and F. Naumann, “Profiling relational data: a survey,” *VLDB J.*, vol. 24, no. 4, pp. 557–581, 2015.
[10] P. Grunwald, “Introducing the Minimum Description Length Principle,” *Adv. Minim. Descr. Lenght Theory Appl.*, no. January 2007, pp. 3–21, 2005.
[11] Y. Li, R. Krishnamurthy, S. Raghavan, S. Vaithyanathan, and H. V Jagadish, “Regular Expression Learning for Information Extraction,” *Proc. Conf. Empir. Methods Nat. Lang. Process.*, no. October, pp. 21–30, 2008.
[12] F. Dwiandriani, “Analisis Dan Perancangan Arsitektur Aplikasi Data Profiling Berbasis Open Source Febri Dwiandriani Program Studi Sistem Informasi Analisis Dan Perancangan Arsitektur Aplikasi Data Profiling Berbasis Open Source Febri Dwiandriani,” p. 127, 2017.