Public Health and Informatics
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210332

Adjustable Data Cleaning Towards Extracting Statistical Information

Argyro MAVROGIORGOU[a,1], Athanasios KIOURTIS[a],
George MANIAS[a] and Dimosthenis KYRIAZIS[a]
[a] *Department of Digital Systems, University of Piraeus, Greece*

Abstract. Each device, organization, or human, is affected by the effects of Big Data. Analysing these vast amounts of data can be considered of vital importance, surrounded by many challenges. To address a portion of these challenges, a Data Cleaning approach is being proposed, designed to filter the non-important data. The functionality of the Data Cleaning is evaluated on top of Global Terrorism Data, to furtherly create policies on how terrorism is affecting national healthcare.

Keywords. Data Cleaning, Health Policies, Global Terrorism Data

1. Introduction

Every device or human, is both producing and surrounded by the term of data, falling under three (3) categories: Fast, Large and Complex data [1], creating numerous challenges. Since Data Cleaning improves the data quality and the analysis outcomes, this paper presents a way of effectively performing cleaning actions through four (4) steps, which are adapting their functionality based on the domain that the cleaning actions must be performed.

For Data Cleaning the authors in [2] are filling the missing values and smoothing out the noise, while in [3] they create materialized views (MVs) that pre-process and avoid complex resource intensive calculations. Furthermore, in [4] they proposed the NADEEF architecture that allows users to specify multiple types of data quality rules.

Regarding the rest of this paper, Section 2 depicts the overall proposed approach, while Section 3 includes the evaluation of the derived results, presenting our concluding remarks.

2. Proposed Approach

The Data Cleaning workflow comprises of four (4) steps. However, the initial basic action of the mechanism is to ingest the data from the various data sources and identify their domain, by discovering and analysing the semantics of the ingested data, and thus the Data Cleaning actions to be performed are adjusted according to the domain's nature. Afterwards, the rest of the services (Fig. 1) initiate their tasks for accomplishing all the actions related to Data Validation, Data Cleaning, Data Verification, and Data Logging.

[1] Corresponding Author, Argyro Mavrogiorgou, 80, M. Karaoli & A. Dimitriou St., 18534 Piraeus, Greece; E-mail: margy@unipi.gr.

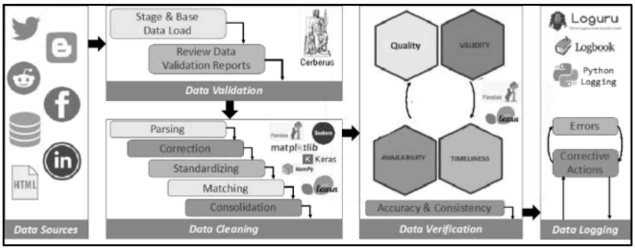
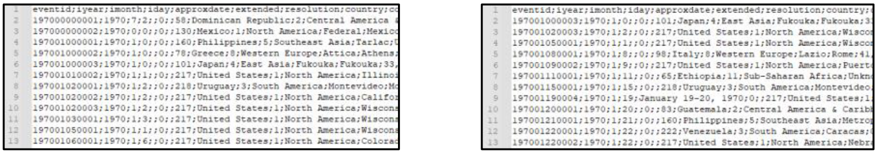


Figure 1. Data Cleaning Workflow.

3. Evaluation and Discussion

To evaluate the Data Cleaning approach, the described data cleaning services have been applied to the Global Terrorism Database (GTD) [5] (Fig. 2), in order to gather insights and policies regarding how global terrorism is affecting the healthcare systems of countries with increased or decreased terrorism attacks accordingly.



```
eventid;iyear;imonth;iday;approxdate;extended;resolution;country;co
197000000001;1970;7;2;;0;;58;Dominican Republic;2;Central America
197000000002;1970;0;0;;0;;130;Mexico;1;North America;Federal;Mexico
197001000001;1970;1;0;;0;;160;Philippines;5;Southeast Asia;Tarlac;
197001000002;1970;1;0;;0;;78;Greece;8;Western Europe;Attica;Athens;
197001000003;1970;1;0;;0;;101;Japan;4;East Asia;Fukouka;Fukouka;33,
197001010002;1970;1;1;;0;;217;United States;1;North America;Illinoi
197001020001;1970;1;2;;0;;218;Uruguay;3;South America;Montevideo;Mo
197001020002;1970;1;2;;0;;217;United States;1;North America;Califor
197001020003;1970;1;2;;0;;217;United States;1;North America;Wiscons
197001030001;1970;1;3;;0;;217;United States;1;North America;Wiscons
197001050001;1970;1;1;;0;;217;United States;1;North America;Wiscons
197001060001;1970;1;6;;0;;217;United States;1;North America;Colorad
```

(a)

```
eventid;iyear;imonth;iday;approxdate;extended;resolution;country;
197001000003;1970;1;0;;0;;101;Japan;4;East Asia;Fukouka;Fukouka;3
197001020003;1970;1;2;;0;;217;United States;1;North America;Wisco
197001050001;1970;1;1;;0;;217;United States;1;North America;Wisco
197001080001;1970;1;8;;0;;98;Italy;8;Western Europe;Lazio;Rome;41,
197001090002;1970;1;9;;0;;217;United States;1;North America;Puert
197001110001;1970;1;11;;0;;65;Ethiopia;11;Sub-Saharan Africa;Unkn
197001150001;1970;1;15;;0;;218;Uruguay;3;South America;Montevideo;
197001190004;1970;1;19;January 19-20, 1970;0;;217;United States;1;
197001200001;1970;1;20;;0;;83;Guatemala;2;Central America & Carib
197001210001;1970;1;21;;0;;160;Philippines;5;Southeast Asia;Metro
197001220001;1970;1;22;;0;;222;Venezuela;3;South America;Caracas;
197001220002;1970;1;22;;0;;217;United States;1;North America;Nebr
```

(b)

Figure 2. (a) Ingested Data, **(b)** Cleaned Data.

Several rows of the GTD were dropped, where, out of the 5,000 different instances, 1,563 instances were dropped after three (3) iterations of the provided mechanism. It is within our next goals to perform additional evaluation, to deploy the services in such a way to run in parallel and, respecting privacy and security concerns [6].

Acknowledgment

The research leading to the results presented in this paper has received funding from the European Union's funded Project PolicyCLOUD under grant agreement no 870675.

References

[1] Oussous A, et al.. Big Data technologies: A survey, Journal of King Saud University-Computer and Information Sciences. 2018;30(4): 431-448.

[2] Somasundaram RS, et al. Evaluation of three Simple Imputation Methods for Enhancing Preprocessing of Data with Missing Values, International Journal of Computer Applications. 2011;21(10).

[3] Saqib M, et al. Improve Data Warehouse Performance by Preprocessing and Avoidance of Complex Resource Intensive Calculations, International Journal of Computer Science Issues. 2012;9(2).

[4] Dallachiesa M, et al. NADEEF: A Commodity Data Cleaning System, ACM SIGMOD International Conference on Management of Data. 2013.

[5] Global Terrorism Database (GTD), http://ghdx.healthdata.org/record/global-terrorism-database

[6] Kiourtis A, et al. Towards a Secure Semantic Knowledge of Healthcare Data Through Structural Ontological Transformations. Joint Conf. on Knowledge-Based Software Engineering. 2018:178-188.