Data Cleansing en entornos Big Data: Mapeo Sistemático de la Literatura

Caffetti Yanina A.[1], Eckert Karina B. [1,2], Ruidias Héctor J.[1,2] and Vera Laceiras María Silvia[1]

[1] Universidad Nacional de Misiones, Misiones, Argentina
[2] Universidad Gastón Dachary, Misiones, Argentina

yanina.caffetti@fcf.unam.edu.ar, karinaeck@gmail.com, chandra149@gmail.com, vlhsilvia@fceqyn.unam.edu.ar

Resumen. La tecnología Big Data tiene por objetivo la gestión de grandes volúmenes de datos e información de manera inteligente que ayude a una correcta toma de decisión. Las etapas del trabajo en Big Data incluyen muchas decisiones que deben ser tomadas por el usuario, para garantizar el éxito del objetivo propuesto, entre ellas la limpieza y pre-procesamiento de datos. El presente artículo es un Mapeo Sistemático de la Literatura, que busca identificar las metodologías, técnicas o herramientas utilizadas para el tratamiento de datos basura (dirty data) o limpieza de datos (data cleansing), en entornos Big Data. Existe, en la literatura actual, cierta escasez de publicaciones específicas, aun siendo un tema de suma relevancia para el éxito de este tipo de proyectos, donde se requiere procesamientos que cumplan con las características propias del entorno Big Data.

Keywords: Data Cleansing, Dirty Data, Big Data, Method, Treatment.

1 Introducción

El crecimiento de los datos es considerado exponencial, algunos autores indican que el volumen de los datos digitales se duplicará cada dos años [1]. Precisamente, a diario se generan grandes volúmenes provenientes de diferentes fuentes y formatos, especialmente en entornos de Big Data (BD), lo que aumenta el desafío de verificar la calidad de estos datos, que pueden ser imprecisos, tener anomalías o no ser adecuados para el análisis o procesamiento afectando así la precisión de los resultados obtenidos [2].

Entre los inconvenientes asociados a la calidad de los datos crudos, se pueden mencionar los siguientes problemas típicos: ausencia de datos, valores ficticios o predeterminados, ruido, datos erróneos, datos inconsistentes, datos crípticos, claves primarias duplicadas, identificadores no únicos, campos multipropósito y violación de reglas comerciales [3]. La limpieza de datos, incluye diferentes técnicas de detección y representa un proceso complejo que requiere mucho tiempo para poder garantizar que los datos limpios tengan una mejor calidad. Este pre-procesamiento es útil y fundamental para garantizar la fase siguiente o de análisis [2], [4]. El proceso de limpieza

de datos se define en cinco fases; (1) análisis de datos, (2) definición de flujo de trabajo de transformación y regla de mapeo, (3) verificación, (4) transformación y (5) reflujo de datos limpios. Dicho procedimiento consiste básicamente en identificar errores y corregirlos, para así obtener resultados que colaboren en el proceso de toma de decisiones. Por estas razones, el proceso de detección y limpieza de anomalías dentro de los datos recopilados se convierte en un factor crítico. Según el Data Warehousing Institute en su web https://tdwi.org, los datos sucios o erróneos conocidos como Dirty Data cuestan más de $600 mil millones por año en negocios de EEUU; los mismos son causados por diversos factores (datos obsoletos, incompletos, duplicados y/o sin formato, etc.) [1], [2], [4], [5].

El objetivo del presente artículo es indagar sobre los procesos de limpieza de datos, metodologías, técnicas y/o herramientas para el tratamiento de datos sucios en el ámbito de BD, mediante un Mapeo Sistemático de la Literatura (MSL) del tema. El artículo está estructurado de la siguiente manera, en la sección 2 se describe la metodología utilizada (MSL), luego en la sección 3 se exponen el análisis de los resultados obtenidos, para finalmente en la sección 4 presentar las conclusiones del trabajo realizado.

2 Mapeo Sistemático de la Literatura

Como primera fase del MSL [6] se definió el problema a través del planteo de la siguiente pregunta de investigación: ¿Cuáles son las metodologías, métodos, técnicas y herramientas utilizadas para la limpieza de datos en entornos de Big Data?

Las subpreguntas de investigación derivadas orientan las fases subsecuentes, desde la búsqueda hasta el análisis de la información; posibilitando la apropiación y producción del tema, se plantearon dos:

SPI1: ¿Qué tipos de estudios hay respecto a la temática?

SPI2: ¿Qué tipo de tratamientos son los más utilizados en la actualidad?

Tabla 1. Fuentes de datos

Fuentes de datos	Sitio Web	Artículos
ACM Digital Library (ACM)	https://dl.acm.org	1712
IEEE Xplore (IEEEX)	https://ieeexplore.ieee.org	34
Springer Link(SL)	https://link.springer.com	129
ScienceDirect (SD)	https://www.sciencedirect.com	599
DSpace MIT (DMIT)	https://dspace.mit.edu/	51
Scopus (SC)	https://www.scopus.com/home.uri	66

Como segunda fase, se estableció la estrategia de búsqueda y el proceso de selección de los trabajos de investigación. Para ello, se determinaron las cadenas de búsqueda con la finalidad de identificar la literatura relevante. Luego de un proceso de refinamiento en la combinación términos y operadores lógicos utilizados para la confección de las cadenas de búsqueda, se seleccionó la siguiente: *"(Data Cleansing OR*

Dirty Data) AND (Methodology OR Method OR Treatment OR Tool) AND Big Data" que se aplicó a las fuentes de datos indicadas en la Tabla 1, y en donde también se refleja la cantidad de artículos obtenidos por cada fuente de datos. Siendo un total de 2591 artículos seleccionados. En cada buscador se aplicaron filtros específicos, tales como los de restringir la búsqueda específicamente a artículos de investigación ("research articles" por ejemplo en el caso de SD) o por campo de conocimiento que permitieran la mayor especificidad posible ("computer science" fue el filtro que se pudo emplear en plataformas tales como SL, SD) con la finalidad de reducir la muestra.

Posteriormente se eliminaron los estudios duplicados, se definieron y aplicaron los criterios de inclusión y exclusión necesarios para realizar un refinamiento de la búsqueda e identificar aquellos que proporcionan relación directa con la pregunta de investigación.

Criterios de Inclusión (CI):

CI1: Documentos publicados en los últimos 5 años.

CI2: Publicaciones realizadas en revistas, libros, conferencias o workshop.

CI3: Trabajos que refieran específicamente a metodologías, métodos, técnicas o herramientas para el tratamiento o limpieza de datos en los títulos y/o resúmenes.

Criterios de Exclusión (CE):

CE1: Publicaciones previas al 2017.

CE2: Trabajo en sitios que no permiten la disponibilidad del texto completo.

3 Análisis y Discusión

Al aplicar los criterios de inclusión (CI) y exclusión (CE) planteados, se redujo considerablemente la cantidad, obteniendo un total de 11 documentos. Se entiende entonces que el marco de referencia acerca de trabajos realizados buscando una normalización en los datos y el tipo de limpieza de los mismos es un campo aún en desarrollo especialmente en entornos de BD; precisamente en la Fig. 1 se puede notar la tendencia existente entre cantidad de artículos divulgados en los últimos 5 años y los tipos de publicaciones detectados (revistas y conferencias, en este estudio).

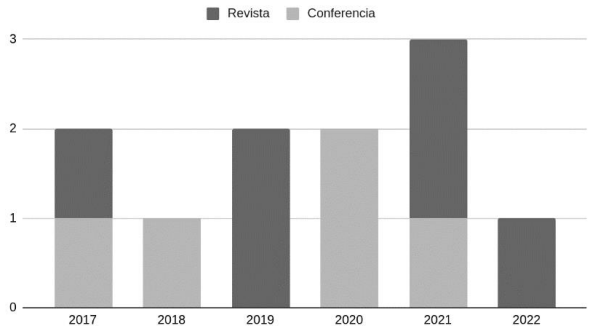

Fig. 1. Relación artículos por tipo y año.

Los artículos resultantes coinciden en que si bien el proceso de limpieza de datos puede ser abordado con herramientas convencionales (tales como hojas de cálculo), en entornos de BD al considerar el volumen de datos y los tiempos de procesamiento requerido, resultan inviables recurrir a tales técnicas tradicionales y se requiere un enfoque sistemático. Además se detecta la necesidad de automatizar y adecuar a los objetivos de los clientes la detección de datos relevantes y ofrecer la confiabilidad necesaria para suponer una mejora en procesos ante un escenario de pérdida de precisión debido al dirty data. Algunas de las técnicas se asientan más en métodos que herramientas, apoyándose fuertemente en la confianza que implica contar con conocimiento experto. Se destacan los métodos basados en técnicas de Inteligencia Artificial, como ser técnicas de aprendizaje automático (machine learning), agrupamiento, especificación y selección de reglas, base de conocimiento y crowdsourcing [2], [3], [4], [7], [8], [9], [10], [11], [12], [13], [14].

4 Conclusiones

En la era de Big Data, la cantidad de datos sigue aumentando, a su vez que se vuelven más complejos debido a su diversidad y combinación (datos estructurados, semiestructurados y no estructurados), lo que hace que la calidad de los mismos disminuya en muchos casos dado que los datos recopilados están sucios (fenómeno conocido como dirty data); donde los procesos tradicionales de limpieza no son adecuados en términos de cantidad, complejidad y velocidad requeridos en proyectos de BD.

En los artículos recuperados y definidos como relevantes, se encontraron ciertas especificaciones en cuanto a metodologías y herramientas en data cleansing. A su vez se identifica la necesidad de investigaciones sobre limpieza de datos centradas en los criterios de BD, donde se cubran las características conocidas como las “5 V” de BD: Volumen, Variedad, Velocidad, Valor y Veracidad. Además de contar con expertos en el área que validen y verifiquen los datos a procesar en las siguientes etapas.

A partir del MSL presentado en esta investigación, se recuperaron artículos relevantes en cuanto a propuestas de tratamiento y limpieza de datos en Big Data, sin embargo no se encuentran del todo estandarizados. En cambio, se han visto como una variación de técnicas más tradicionales como ETL (Extract, Transform and Load) son utilizados. Queda como futura línea de investigación, el planteo de formular un estándar para la aplicación de la data cleansing en entornos BD.

Referencias

[1] M. I. Hossen, M. Goh, A. Hossen, and Md. A. Rahman, “A Study on the Aspects of Quality of Big Data on Online Business and Recent Tools and Trends Towards Cleaning Dirty Data,” in *2020 11th IEEE Control and System Graduate Research Colloquium (ICSGRC)*, Aug. 2020, pp. 209–213. doi: 10.1109/ICSGRC49013.2020.9232648.

[2] F. Ridzuan and W. M. N. Wan Zainon, “A Review on Data Cleansing Methods for Big Data,” *Procedia Computer Science*, vol. 161, pp. 731–738, Jan. 2019, doi: 10.1016/j.procs.2019.11.177.

[3] K. Stöger, D. Schneeberger, P. Kieseberg, and A. Holzinger, “Legal aspects of data cleansing in medical AI,” *Computer Law & Security Review*, vol. 42, p. 105587, Sep. 2021, doi: 10.1016/j.clsr.2021.105587.

[4] T. K. Dang, D. K. Nguyen, and L. M. Tuan, “OpenK: An Elastic Data Cleansing System with A Clustering-based Data Anomaly Detection Approach,” in *2021 15th International Conference on Advanced Computing and Applications (ACOMP)*, Nov. 2021, pp. 120–127. doi: 10.1109/ACOMP53746.2021.00023.

[5] Z. Opršal and J. Harmáček, “Clean aid or dirty aid? The environmentalization of Czech foreign aid.,” *Journal of Cleaner Production*, vol. 224, pp. 167–174, Jul. 2019, doi: 10.1016/j.jclepro.2019.03.198.

[6] B. A. Kitchenham, D. Budgen, and O. Pearl Brereton, “Using mapping studies as the basis for further research – A participant-observer case study,” *Information and Software Technology*, vol. 53, no. 6, pp. 638–651, Jun. 2011, doi: 10.1016/j.infsof.2010.12.011.

[7] F. Ridzuan and W. M. N. Wan Zainon, “Diagnostic analysis for outlier detection in big data analytics,” *Procedia Computer Science*, vol. 197, pp. 685–692, Jan. 2022, doi: 10.1016/j.procs.2021.12.189.

[8] H. A. Sulistyo, T. F. Kusumasari, and E. N. Alam, “Implementation of Data Cleansing Null Method for Data Quality Management Dashboard using Pentaho Data Integration,” in *2020 3rd International Conference on Information and Communications Technology (ICOIACT)*, Nov. 2020, pp. 12–16. doi: 10.1109/ICOIACT50329.2020.9332030.

[9] P. Petrova, V. Jotsov, and V. Sgurev, “Puzzle Methods for Automatic Selection of Data Cleansing Techniques,” in *2018 International Conference on Intelligent Systems (IS)*, Funchal - Madeira, Portugal, Sep. 2018, pp. 820–826. doi: 10.1109/IS.2018.8710580.

[10] Y. Lei, X. Zhou, X. Xu, and F. Jia, “A dirty data recognition method for machinery condition monitoring in big data era,” in *IECON 2017 - 43rd Annual Conference of the IEEE Industrial Electronics Society*, Beijing, China, Oct. 2017, pp. 7061–7066. doi: 10.1109/IECON.2017.8217235.

[11] J. G. Lawson and D. A. Street, “Detecting dirty data using SQL: Rigorous house insurance case,” *Journal of Accounting Education*, vol. 55, p. 100714, Jun. 2021, doi: 10.1016/j.jaccedu.2021.100714.

[12] W. Fan and C. Hu, “Big Graph Analyses: From Queries to Dependencies and Association Rules,” *Data Sci. Eng.*, vol. 2, no. 1, pp. 36–55, Mar. 2017, doi: 10.1007/s41019-016-0025-x.

[13] D. C. Setyawan, T. F. Kusumasari, and E. N. Alam, “Data Cleansing Processing using Pentaho Data Integration: Case Study Data Deduplication,” in *2020 6th International Conference on Science and Technology (ICST)*, Sep. 2020, vol. 1, pp. 01–05. doi: 10.1109/ICST50505.2020.9732824.

[14] S. Salloum, J. Z. Huang, and Y. He, “Exploring and cleaning big data with random sample data blocks,” *J Big Data*, vol. 6, no. 1, p. 45, Jun. 2019, doi: 10.1186/s40537-019-0205-4.