Big Data Preprocessing, Techniques, Integration, Transformation, Normalisation, Cleaning, Discretization, and Binning ⊗

Pranali Dhawas (/affiliate/pranali-dhawas/458779/), Abhishek Dhore (/affiliate/abhishek-dhore/458780/), Dhananjay Bhagat (/affiliate/dhananjay-bhagat/458781/), Ritu Dorlikar Pawar (/affiliate/ritu-dorlikar-pawar/458782/), Ashwini Kukade (/affiliate/ashwini-kukade/458783/), Kamlesh Kalbande (/affiliate/kamlesh-kalbande/458784/)

Source Title: Big Data Analytics Techniques for Market Intelligence (/gateway/book/322396)

Pages: 24

ISBN13: 9798369304136ISBN13 Softcover: 9798369304143EISBN13: 9798369304150

DOI: 10.4018/979-8-3693-0413-6.ch006

Abstract

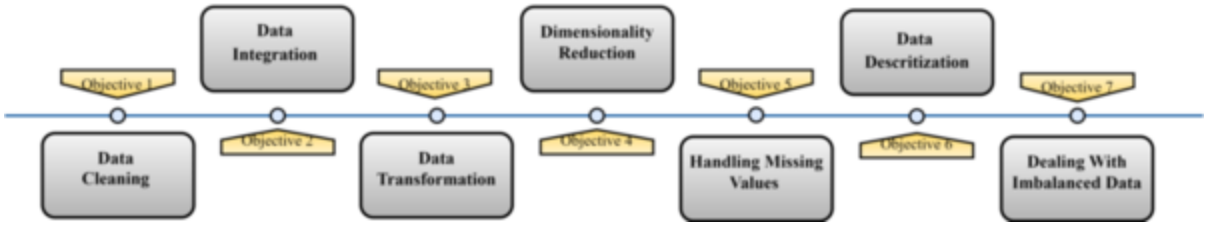
“Unleashing the Power of Big Data: Innovative Approaches to Preprocessing for Enhanced Analytics” is a groundbreaking chapter that explores the pivotal role of preprocessing in big data analytics. It introduces diverse techniques to transform raw, unstructured data into a clean, analyzable format, addressing the challenges posed by data volume, velocity, and variety. The chapter emphasizes the significance of preprocessing for accurate outcomes, covers advanced data cleaning, integration, and transformation techniques, and discusses real-time data preprocessing, emerging technologies, and future directions. This chapter is a comprehensive resource for researchers and practitioners, enabling them to enhance data analytics and derive valuable insights from big data.

1. Introduction To Big Data Preprocessing

Big data preprocessing plays a critical role in the data analysis process by converting raw and unprocessed data into a structured and clean format suitable for analysis. As the volume, velocity, and variety of data continue to grow exponentially, preprocessing becomes increasingly vital for extracting valuable insights and knowledge from large datasets.

The process of big data preprocessing involves employing various techniques and operations to enhance data quality, reduce noise and inconsistencies, handle missing values, and prepare the data for subsequent analysis tasks as shown in figure 1. It significantly contributes to improving the efficiency, accuracy, and effectiveness of data analysis (O. Çelik, 2019).

Figure 1. Objectives of big data preprocessing



(https://igiprodst.blob.core.windows.net:443/source-content/9798369304136_322396/979-8-3693-0413-6.ch006.f01.png?sv=2015-12-11&sr=c&sig=goVJA1ISO%2FYUsT5afpoAe9yu5eJLf%2FHXTpstHuQofeQ%3D&se=2024-11-05T11%3A20%3A49Z&sp=r)

The main objectives of big data preprocessing include:

Data Cleaning: Raw data often contains errors, outliers, duplicates, or inconsistencies. Data cleaning aims to identify and rectify these issues to ensure high data quality. By eliminating noise and irregularities, the resulting clean data provides a reliable foundation for analysis.

Data Integration: Big data originates from diverse sources such as databases, sensors, social media, or IoT devices. Data integration involves combining data from different sources and formats into a unified representation. This step ensures data consistency and compatibility for analysis (Z. Cai-Ming, 2020).

Data Transformation: Data transformation techniques are applied to convert data into a suitable format for analysis. This may involve scaling numerical data, normalizing values, encoding categorical variables, or deriving new features through mathematical or statistical operations. Transformation facilitates data standardization and simplifies subsequent analysis tasks.

Dimensionality Reduction: Dealing with high-dimensional data can pose computational challenges and introduce noise or overfitting problems. Dimensionality reduction techniques help decrease the number of variables or features while preserving crucial information. This simplifies the analysis process and improves computational efficiency (H. S. Obaid, 2019).

Handling Missing Values: Missing data is a common issue in large datasets. Preprocessing techniques include imputing missing values using statistical methods or leveraging imputation algorithms to fill in the gaps. Proper handling of missing data ensures that the analysis is not compromised by incomplete information (T. A. Alghamdi, 2022).

Data Discretization: Discretization involves converting continuous data into categorical or discrete representations. This technique simplifies analysis by reducing the complexity associated with continuous variables. It allows for the application of methods specifically designed for categorical data (P. Gao, 2020).

Dealing with Imbalanced Data: Imbalanced data refers to situations where one class or category is significantly more prevalent than others. Preprocessing techniques address this imbalance by employing methods such as oversampling, under sampling, or generating synthetic samples to achieve a balanced representation of the data.

Big data preprocessing is indispensable for extracting valuable insights from complex datasets. By effectively cleaning, transforming, and organizing the data, preprocessing ensures that subsequent analysis tasks are more accurate, efficient, and reliable. The specific techniques utilized may vary based on the data's nature, analysis objectives, and the challenges posed by the dataset at hand.

Continue Reading (/gateway/chapter/full-text-html/336349)

References

Follow Reference Alghamdi T. A. Javaid N. (2022). A survey of preprocessing methods used for analysis of big data originated from smart grids.IEEE Access : Practical Innovations, Open Solutions, 10, 29149–29171. 10.1109/ACCESS.2022.3157941

Follow Reference Barse S. Bhagat D. Dhawale K. Solanke Y. Kurve D. (2023). Cyber-Trolling Detection System. Available at SSRN4340372.

Follow Reference Çelik, O., Hasanbaşoğlu, M., Aktaş, M. S., Kalıpsız, O., & Kanli, A. N. (2019, September). Implementation of data preprocessing techniques on distributed big data platforms. In *2019 4th International Conference on Computer Science and Engineering (UBMK)* (pp. 73-78). IEEE. 10.1109/UBMK.2019.8907230

Follow Reference Desai V. Dinesha H. A. (2020, November). A hybrid approach to data pre-processing methods. In 2020 IEEE International Conference for Innovation in Technology (INOCON) (pp. 1-4). IEEE.10.1109/INOCON50539.2020.9298378

Hande, T., Dhawas, P., Kakirwar, B., & Gupta, A. (2023, August) Yoga Postures Correction and Estimation using Open CV and VGG 19 Architecture. In *2023 International Journal of Innovative Science and Research Technology*.

Follow Reference Krishna, G. S., Supriya, K., & Rao, K. M. (2022, September). Selection of data preprocessing techniques and its emergence towards machine learning algorithms using hpi dataset. In 2022 IEEE Global Conference on Computing, Power and Communication Technologies (GlobConPT) (pp. 1-6). IEEE. 10.1109/GlobConPT57482.2022.9938255

Gao, P., Han, Z., & Wan, F. (2020, October). Big Data Processing and Application Research. In 2020 2nd International Conference on Artificial Intelligence and Advanced Manufacture (AIAM) (pp. 125-128). IEEE.

Cai-Ming, Z., & Hao-Nan, C. (2020, December). Preprocessing method of structured big data in human resource archives database. In 2020 IEEE International Conference on Industrial Application of Artificial Intelligence (IAAI) (pp. 379-384). IEEE.

Follow Reference Gawhade, R., Bohara, L. R., Mathew, J., & Bari, P. (2022, March). Computerized Data-Preprocessing To Improve Data Quality. In *2022 Second International Conference on Power, Control and Computing Technologies (ICPC2T)* (pp. 1-6). IEEE. 10.1109/GlobConPT57482.2022.9938255

Follow Reference Mo, R., Liu, J., Yu, W., Jiang, F., Gu, X., Zhao, X., & Peng, J. (2019, August). A differential privacy-based protecting data preprocessing method for big data mining. In *2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)* (pp. 693-699). IEEE. 10.1109/TrustCom/BigDataSE.2019.00098

Follow Reference Obaid, H. S., Dheyab, S. A., & Sabry, S. S. (2019, March). The impact of data pre-processing techniques and dimensionality reduction on the accuracy of machine learning. In *2019 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference (IEMECON)* (pp. 279-283). IEEE. 10.1109/IEMECONX.2019.8877011

Follow Reference

Pradha, S., Halgamuge, M. N., & Vinh, N. T. Q. (2019, October). Effective text data preprocessing technique for sentiment analysis in social media data. In *2019 11th international conference on knowledge and systems engineering (KSE)* (pp. 1-8). IEEE. 10.1109/KSE.2019.8919368