


Enhancing Data Quality at ETL Stage of Data Warehousing

Neha Gupta, Manav Rachna International Institute of Research and Studies, Faridabad, India

 <https://orcid.org/0000-0003-0905-5457>

Sakshi Jolly, Manav Rachna International Institute of Research and Studies, Faridabad, India

ABSTRACT

Data usually comes into data warehouses from multiple sources having different formats and are specifically categorized into three groups (i.e., structured, semi-structured, and unstructured). Various data mining technologies are used to collect, refine, and analyze the data which further leads to the problem of data quality management. Data purgation occurs when the data is subject to ETL methodology in order to maintain and improve the data quality. The data may contain unnecessary information and may have inappropriate symbols which can be defined as dummy values, cryptic values, or missing values. The present work has improved the expectation-maximization algorithm with dot product to handle cryptic data, DBSCAN method with Gower metrics to ensure dummy values, Wards algorithm with Minkowski distance to improve the results of contradicting data and K-means algorithm along with Euclidean distance metrics to handle missing values in a dataset. These distance metrics have improved the data quality and also helped in providing consistent data to be loaded into a data warehouse.

KEYWORDS

Data Mining, Data Purgation (DP), Data Quality (DQ), Data Warehouse (DW), EM Modelling, Euclidean Distance, Extract, Minskovi Distance, Transform and Load (ETL)

1. INTRODUCTION

1.1 Data Quality at ETL in Data Warehouse

Information warehousing is a network of decision support tools focused on supporting the analyst's knowledge to make quicker and better decisions. Data warehouse is the compilation of non-volatile, subject-oriented and organized data. The purposes of data warehouse involve:

1. Data Extraction – It is used to collect information from various heterogeneous resources.
2. Data Cleaning - It is used to spot and correct the flaws in the collected data.
3. Data Transformation – It converts the information from legal presentation to warehouse format.
4. Data Loading – It includes arranging, associating, evaluating, partitioning, checking integrities, and developing entities.
5. Refreshing – It is the method of updating the data sources to the warehouse.

DOI: 10.4018/IJDWM.2021010105

Copyright © 2021, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

In many domains, the data warehouse information can be studied and referred for multiple purposes. It is used by rearranging the products and handling the product portfolios relating to the sales and profit of the year to tune production strategies. The information is used for market research by analyzing the customer's interest, their buying time, budget cycle, etc. This analysis plays a vital role in supporting the management of customer relationships in order to make changes according to demand and conditions.

As discussed above, data is the main fuel for any kind of prediction and any kind of operation as well. The initial requirement is to understand the data warehouse components and the data warehouse framework along with the operations which can be done with the data. The information collected from the various resources could be historical information or the accumulated information called as data warehouse. There are various security and quality issues with the data collected from various sources. Data will be collected from various repositories having different formats for the same type of data. To understand the concept, various data formats available in data warehouses have been explained as follows:

1.1.1. Structured Data

This can be historical data that can be compared with database information. The data in the databases can be stored in the form of rows and columns and can be manipulated in the form of rows and columns with simple queries. The queries must target the variables of particular row and column and with simple basic operations on the databases, so that the data can be understood. Most representations of data will be text format and other formats of the database.

1.1.2. Unstructured Data

This type of data usually has multimedia content. The multimedia consists of different formats of images, videos and audios. Sometimes we need to consider the concept of live streaming. Live streaming is the data that is captured using Apache spark as the main base. It consists of components that can handle live streaming information. It can be the main source of operations on big data, cloud computing and can also be the source of data management.

1.1.3. Semi – Structured Data

This format ensures connectivity to the web application. Every web application has some sort of data transfer mechanism and a framework that can hold the data from the user to the database. The information will be carried by those frameworks to the database. In this scenario, data quality cannot be manipulated as there are different technologies that can be used for the information transfer from the client.

A decision making and the implementation of the predictions can be done with the help of valuable information. This kind of information should be with valid quality metrics and the metrics need to be followed to maintain the great accuracy of the data.

Extract, Transformation and Load (ETL) is the process of handling the data quality with the data warehouse and the process of data quality will be affected when in process of pre-processing.

The quality of data can be slightly compromised based on its functions, such as extraction, transformation, cleaning, and loading. Data is affected by several processes depending upon its environment. Even though, after cleaning and filling, there may be residual dirty data, which should be reported and these remaining dirty data can be the reason for failure during the process of data cleaning.

16 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/article/enhancing-data-quality-at-etl-stage-of-data-warehousing/272019

Related Content

The Impact of Big Data on Security

Mohammad Alaa Hussain Al-Hamami (2016). *Big Data: Concepts, Methodologies, Tools, and Applications* (pp. 1495-1518).

www.irma-international.org/chapter/the-impact-of-big-data-on-security/150227

A Method for Generating Comparison Tables From the Semantic Web

Arnaud Giacometti, Béatrice Markhoffand Arnaud Soulet (2022). *International Journal of Data Warehousing and Mining* (pp. 1-20).

www.irma-international.org/article/a-method-for-generating-comparison-tables-from-the-semantic-web/298008

The Role of Schema and Document Matchings in XML Source Clustering

Pasquale De Meo, Giacomo Fiumara, Antonino Noceraand Domenico Ursino (2012). *XML Data Mining: Models, Methods, and Applications* (pp. 125-153).

www.irma-international.org/chapter/role-schema-document-matchings-xml/60907

P2P-COVID-GAN: Classification and Segmentation of COVID-19 Lung Infections From CT Images Using GAN

Nandhini Abirami, Durai Raj Vincentand Seifedine Kadry (2021). *International Journal of Data Warehousing and Mining* (pp. 101-118).

www.irma-international.org/article/p2p-covid-gan/290272

Filter-Wrapper Incremental Algorithms for Finding Reduct in Incomplete Decision Systems When Adding and Deleting an Attribute Set

Nguyen Long Giang, Le Hoang Son, Nguyen Anh Tuan, Tran Thi Ngan, Nguyen Nhu Sonand Nguyen Truong Thang (2021). *International Journal of Data Warehousing and Mining* (pp. 39-62).

www.irma-international.org/article/filter-wrapper-incremental-algorithms-for-finding-reduct-in-incomplete-decision-systems-when-adding-and-deleting-an-attribute-set/276764