UNIVERSITY OF SCIENCE AND TECHNOLOGY

COLLEGE OF GRADUATE STUDIES AND ACADEMIC ADVANCEMENT

Faculty of Computer Science and Information Technology

Design Preprocessing Tool for Structure Data
(Detection and Handling Outliers)

By

Abobaker Yousif Adam Mohammed

A Thesis

Submitted to the College of Graduate Studies and Academic Advancements
In Partial Fulfillment of the Requirement for the Degree of Master of Science in Information Technology

Supervisor

Dr. Atif Ali Mohamed

September 2016

Abstract

Data preprocessing describes any type of processing performed on raw data to prepare it for another processing procedure. Commonly used as preliminary data mining practice, data preprocessing transforms the data into a format that will be more easily and effectively processed for the purpose of the user.

Data preprocessing an important issue for data warehousing and data mining real world data tend to be incomplete, noisy and inconsistent includes –data. Then there are needs for design preprocessing tools to find out solutions for these issues.

The objective of this research is to design a preprocessing tool for structure data and the sub objectives detect and handle outlier's data to apply this process in structure data repository. Java is used for designing the tool.

The tool was tested and the result shows it is effectiveness preprocess data from structure data repositories

المستخلص

تحضير البيانات تصف أي نوع من المعالجات التي تجرى على البيانات الاولية لتجهيزها لمعالجات أخرى. عادة ما تستخدم كمرحلة اولية في عملية تنقيب البيانات، حيث يتم في تحضير البيانات تحول البيانات الي هئية تكون معالجتها اكثر سهولة وفعالية للاغراض التي يريدها المستخدم .

نجد ان تحضير البيانات من القضايا المهمة في لمستودع البيانات وتنقيب البيانات حيث ان في العالم الخارجي تميل البيانات إلى أن تكون بيانات غير تامه، مزعجه وغير متناسقة ومتضاربه .لذلك اظهرت الحوجه الي تصميم اداه تعمل علي لتحضير ومعالجة البيانات وإيجاد حلول لهذه القضايا .يهدف هذا البحث لتصميم أداة لتحضير كميه ضخمه من البيانات وذلك باستخام لغه الحافا ، للوصول . تم اختبار الاداه واظهرت النتائج فاعلية تحضير البيانات علي هيكلية مستودع البيانات.

1.1 Introduction

Data preprocessing concept seeks to streamline and improve the quality of data hence making it more reliable. Data preprocessing does this by removing the extraneous information and mining the key features of the data to simplify the pattern detection process difficulties without disregarding any critical information[1].

Data mining pre processing means preparing data .It is the one of the important and compulsory tasks an important issue for both data warehouse and dada mining. Before applying the data mining techniques like association, classification or clustering noisy And Outliers Should Be Removed.

The data “cleaning” routine entails various tasks such as; data acquisition, filling missing data values, unifying date formats, conversion of nominal values to numeric data, identification of outliers and smoothening of noisy data, and correcting inconsistent data[2].

Data integration involves combining data residing in different sources and providing users with a unified view of these data. This process becomes significant in a variety of situation, which include both commercial (when two similar companies need to merge their data bases) and scientific (combining research results from different bioinformatics repositories) domains[3].

1.2 Research Problem

Structure data has many challenges that face organizations when they need to extract information or knowledge. This challenge includes errors, lack of data quality and noise. Then there is need for preprocessing tools to find out solutions for these challenges.

Hat everyone knows about the outlier’s data, but most people aren’t sure how to deal with. Operation of outliers detects and handle.

1.3 Research Objective

The main objective of this research is to design preprocessing tool (AMAS) for structure data that clean data from structure repository with respective to detect and handle missing data, outliers, instant duplicate and integrate the data from heterogonous resource.

The sub objective is to implement detection and handling outliers' data by used k means algorithm in java programming language**.**

1.4 Research Methodology

For design preprocessing tool (AMAS) there are many algorithms used such as a K-means to detect and handle noise data, linear search to detect missing values, Proposed Algorithm for Integration process, and Proposed Algorithm for Duplicate Detecting, while a Java programming language is used for implementation.

1.5 Research Structure

This research is divide into five chapters where, chapter two discuss the main concept of data mining and related work for previous researches, chapter three contain the technique and Algorithms that used in this research, chapter four explain the methodology and implantation of preprocessing tool (AMAS) , chapter five Represent the conclusion and recommendation .