|  |
| --- |
| **A Project Based Seminar Report**  **on**  **‘Purging and cleaning of impure datasets’**   Submitted to theSavitribai Phule Pune UniversityIn partial fulfillment for the award of the Degree ofBachelor of EngineeringinInformation TechnologybySatwik Suhas Ramchandre(T150028563 / 306074 & TE - IT)T.E. (Information Technology)Under the guidance of**Mrs. Himangi Pande** **MIT_Logo**  **DEPARTMENT OF INFORMATION TECHNOLOGY**  **MAHARASHTRA INSTITUTE OF TECHNOLOGY**  **PUNE-411038**  **Academic Year 2018-2019**  **Semester II**  Affiliated to  uop_logo  Savitribai Phule Pune University |
| MIT_Logo  **MAEER’S**  **MAHARASHTRA INSTITUTE OF TECHNOLOGY, PUNE** DEPARTMENT OF INFORMATION TECHNOLOGY**CERTIFICATE** This is to certify that the project based seminar report entitled**”** **Purging and cleaning of impure datasets”** being submitted by **Shardul Pharande (T150028571/306071 TE - IT)** is a record of bonafide work carried out by him/her under the supervision and guidance of Prof. Himangi Pande in partial fulfillment of the requirement for **TE (Information Technology Engineering) a 2015 course** of Savitribai Phule Pune University, Pune in the academic year 2018- 2019.  Date: 05/04/2019  Place: Pune Prof. Sumedha Sirsikar **Seminar Guide Head of the Department, IT**  Dr. L.K. Kshirsagar  Principal  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  This Project Based Seminar report has been examined by us as per the Savitribai Phule Pune University, Pune requirements at Maharashtra Institute of Technology, Pune 411038 on . . . . . . . . . . .  Internal Examiner External Examiner | | |

###### *ACKNOWLEDGEMENT*

I take this opportunity to express my sincere appreciation for the cooperation given by Prof. Sumedha Sirsikar, HOD (Dept. of Information Technology) and need a special mention for all the motivation and support.

I am deeply indebted to my guide Prof. Himangi Pande for completion of this seminar for which she has guided and helped me going out of the way.

For all the efforts behind the seminar report, I would also like to express my sincere appreciation to the staff of the Dept. of Information Technology, MIT Pune, for their extended help and suggestions at every stage.

**Satwik Ramchandre**

Roll No: 306074

T.E. (Information Technology)

MIT, Pune 411038

## Abstract

Data mining provides an user oriented approach to extract and analyze hidden patterns from a massive database.

**Keywords:**

*Data mining, Data collection, Data cleaning, Algorithm*

## INDEX

|  |  |  |  |
| --- | --- | --- | --- |
| **Sr. Topic**  **No** | | | **Page No.** |
|  |  |  |  |
| 1. | INTRODUCTION | |  |
|  | 1.1 | Introduction to the project |  |
|  |  |  |  |
|  | 1.2 | Motivation for the project |  |
|  | 1.3  1.4 | Aim and Objective  Introduction to the seminar topic |  |
|  |  | |  |
| 2. | LITERATURE SURVEY | |  |
|  | 2.1 | Crowd-sourced Data Collection of Physical Activity and Health Status: An App Solution |  |
|  |  |  |  |
|  | 2.2  2.3  2.4  2.5 | Data Cleaning for Data Quality  Cleaning Framework for BigData: An Interactive approach for data cleaning  Data Cleaning – A thorough analysis and survey on Unstructured data  An Efficient Data Cleaning Algorithm based on Attributes Selection |  |
|  |  |  |  |
|  |  |  |  |
| 3.  4. | DATA CLEANING  3.1  3.2  ALGORITHMS……  4.1  4.2  4.3  4.4  4.5 | Why Data cleaning?  Importance of study of Data Cleaning and its algorithms  Variegated Data Swabbing  Association and Repairing Algorithm  Missing values replaced with leaf node values  Approximate duplicate record cleaning Algorithm  Normalization Algorithm |  |
| 5. | INFERENCE |  |  |
| 6.  7. | CONCLUSION  REFERENCES | |  |
|  |  | |  |
|  |  | |  |

I]List of Figures

II]List of Tables

**CHAPTER 1**

1. **INTRODUCTION TO PROJECT TOPIC**

**1.1 Introduction to Project**

Students are expected to write brief introduction of the project topic. This section may be common for all the students of that group. However students can have different approach in explaining their project.

Since we are not going to get separate project report from the students, it is good to have one or two pages common for all the students of that project. Once this project introduction and aim objectives of explained, students can start with actual seminar content writing.

**1.2 Motivation behind project topic**

Students are expected to write” Why they thought about this project”. They could also explain the” Explain need of the project”. Any statement which motivated to take up this project

**1.3 Aim and Objective(s) of the work**

Project aims are a statement of the overall ” Why” of the project. A project aim is a statement starting with the words: ” The aim of this project is. . . ”

That statement tells the reader what your overall goal is, what it is you want to achieve. It does not go into details or describe specific tasks.

Project objectives tell the ” How” of the project aim. You want to achieve your goal and the objectives state how this will be done - what ma- jor tasks will be undertaken and what your major targets are. Most projects will have between six and nine objectives. A project objective is usually a statement starting with the words ” To. . . ”

Objectives are subsidiary to aims and are the steps you are going to take to answer your problem statement or a specific list of tasks needed to accomplish the goals of the project. This must be highly focused and feasible and should address the more immediate project outcomes

**1.4 Introduction to Seminar Topic**

Internal guides need to clearly identify separate topics for individual students while preparing seminars. It should be based on project topic/area.

**CHAPTER 2**

**2. LITERATURE SURVEY OF ‘Purging and cleaning of impure datasets’**

**2.1 Crowdsourced Data Collection of Physical Activity and Health Status: An App Solution :** **Daniel Kelly, Brian Caulfield, and Kevin Curran**

**Publication:** ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2017 P. Perego et al. (Eds.): MobiHealth 2016, LNICST 192, pp. 151–159, 2017

Health status and related data and metrics are vital to understand a patient’s health. However , the precision and the means of measuring such data are very limited. In this paper , an app system for mobile phones is implemented to take into account the problem of lack of such health related data , and is gathered using sensors and a SF-36 questionnaire. Preliminary analysis of the data also shows a statistically significant correlation between the amount of time a participant is active and the health status of the participant.

**2.2 Data Cleaning for Data Quality**

**S. Swapna, P. Niranjan, B. Srinivas, R. Swapna [2]**

**Publication:** [2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)](https://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=7589474)

**C**onsistent data is essential and reliable for decision making. The author(s) proposes an algorithm which constructs a decision tree for every attribute, and all the missing values can be replaced by the leaf node values. Thus there would be a consistency in the database as missing values will be dealt with efficiently. Cleaning is done on the dirty data, mainly those with:

* Missing Values
* Dummy Values
* Cryptic Data
* Contradicting Data

**2.3 Cleaning Framework for BigData: An Interactive approach for data cleaning**

**Hong Liu, Ashwin Kumar TK, Johnson P Thomas, Xiaofei Hou[3]**

**Publication:** [2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService)](https://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=7474310)

The core of the system proposed by the author(s) is data association and data repairing. Association is the process of identifying the similarity between objects and then link them together. Repairing, as the name suggests, is the actual process of making the database reliable by fixing errors in the data.

The research done in this paper shows that data association can effectively help with data repairing. The paper unifies these two steps into a single framework based on context patterns, usage patterns, metadata, and repairing rules.

**2.4 Data Cleaning – A thorough analysis and survey on**

**Unstructured data**

**Virender Kumar, Cherry Khosla[4]**

**Publication:** [2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence)](https://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=8421099)

In this paper, the author(s) explains the importance of data cleaning by taking an example of pollution dataset. The main focus of this paper is:

* Analysis and visualization of unstructured data
* Survey on methods to clean dirty data.

**2.5 An Efficient Data Cleaning Algorithm based on Attributes Selection**

**Ling He, Zhongnan Zhang1, Yize Tan, Minghong Liao[5]**

**Publication**: [2011 6th International Conference on Computer Sciences and Convergence Information Technology (ICCIT)](https://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=6302083)

Detection of the duplicity of 2 records amongst an extremely large dataset is a tedious arithmetic process. Finding out approximate duplicate records is an important task for such data in the data cleaning the process. This paper proposes an algorithm of approximately duplicate records cleaning based on attribute selection after analyzing the existing basic sorted neighborhood method (SNM) and multi-pass sorted neighborhood method (MPN).

**2.6 Distance-based Pattern Matching of DNA Sequences for Evaluating Primary Mutation**

**Berlian Al Kindhi, Muhammad Afif Hendrawan, Diana Purwitasari, Tri Arief Sardjono, Mauridhi Hery Purnomo[6]**

**Publication**: 2017 IEEE 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)

String matching is the usual method to find out DNA Pattern, but basic string matching methods are unable to recognize the mutations of viruses and bacteria (for example, HCV – i.e. Hepatitis C Virus infection). In this paper, the hamming method is used, which can accept character mismatch. The lengths of HCV primary characteristics (the characters that cannot be changed, deleted or added in DNA sequence) vary, and thus the hamming count is unbalanced. This paper proposes to normalize the primary before being tested, which would result in the balanced hamming count. In this paper, the research question is to recognize a mutated sequence by approximating similarities of isolated data and primary sequences with string matching method.

**CHAPTER 3**

**3. DATA CLEANING**

**Data cleaning** is the process of detecting and correcting (or removing) corrupt or inaccurate [records](https://en.wikipedia.org/wiki/Storage_record) from a record set, [table](https://en.wikipedia.org/wiki/Table_(database)), or [database](https://en.wikipedia.org/wiki/Database) and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the [dirty](https://en.wikipedia.org/wiki/Dirty_data), unclean or coarse data.

**3.1 Why Data cleaning?**

In today’s world, the production of data is at an exponentially high scale. If the data is not cleaned, about 25% of the data in the database is lost or is made unusable in a year.

Even the best analytics in the world are going to fail when it comes to analyzing extremely unclean data. Endless false alarms about heart attacks that don't happen or, even worse, a heart attack that kills someone, which came without warning, because of faulty data failing to alert anyone to an imminent problem, would clearly be a disaster.

Thus, it is extremely important that the database should be cleaned and maintained frequently.

**3.2 Importance of study of Data Cleaning and its algorithms**

Importance of data cleaning:

* Improves decision making process
* Increases productivity
* Eases the retrieval of data, further making the job efficient

Importance of studying data cleaning algorithms:

Data cleaning can be complicated, time consuming, and expensive, but it is a necessary step in any data-related system since poor-quality data may not be suitable to achieve the intended purposes. Aim is to study an efficient system that uses data cleaning algorithms to lower the effort as well as cost which is put in.

**CHAPTER 4**

**4.** **ALGORITHMS**

The types of algorithms studied are:

1. Variegated Data Swabbing
2. Association and repairing algorithm
3. Replacement by leaf node values
4. Approximate duplicate record cleaning algorithm
5. Normalization (Unconventional)

**4.1 Variegated Data Swabbing Algorithm**



***Fig. 1. Variegated Data Swabbing Architecture***

This algorithm consists of 4 phases, namely:

* Phase 1: Integration of heterogeneous Sources
* Phase 2: Replacing Missing Values
* Phase 3: De-duplication
* Phase 4: Misspelling detection
* Phase 5: Loading

In the **first phase**, data is collected and stored in tables from different sources but the same domain. These separate datasets are then combined to form a single dataset. Data profiling operation is further carried out, which examines the data available and collects statistics or informative summaries about that data, which might be used for other future purposes.

Following the first phase is the **second phase**, which analyzes the data which results in identification of missing values in the integrated dataset which are filled by “NaN”.

Further in the **third phase**, the duplicated values are removed and this process is called de-duplication.

In the fourth phase, it takes dataset with no missing values and duplications. Its output is a dataset containing the misspellings with the correct suggestions.

Finally, in the **fifth phase** of this framework, the dataset free from missing and redundant values is loaded back into the original dataset.

Algorithm **–**

1. Load the dataset A and B
2. Integrate the two datasets A and B, and name the new dataset as D.

Σ A + B = D

1. Eliminate the redundant rows of data from the new dataset D to form a new dataset D’.

D – D1 = D’

1. Fetch the missing values in each column of the dataset D.
2. Replace each missing value with the nominal value of the column,

If the data in column is numeric,

if(column datatype == numeric)

NaN = (col data1 + col data2 + … + col data n)/n

Where col data = column data

1. Check for the columns of dataset which contains string type data

if (column datatype == string)

Check column, value from dictionary

if (value = valid)

{

Show suggestions

}

else

Return false

**4.2 Association and Repairing Algorithm**

This algorithm focuses on 3 types of errors:

* Value Error – Actual value is wrong (Eg. 39 instead of 93)
* Conditional Error – Logically correct but wrongly depicted (Eg. 09 is a wrong year, it should be depicted as 2009)
* Missing Values - The concept of missing value refers to no data value is stored for the attribute in a tuple of dataset.

**CHAPTER 5**

**5. INFERENCE**

**CHAPTER 6**

**6. CONCLUSION**

Conclusions usually serve two functions. The fi is to summarize and bring together the main areas covered in the writing, which might be called ”look- ing back” . The second is to give a fi comment of your seminar.

For example your could say this seminar is undertaken to explain...... and evaluate ..... . This study has found that generally ....... . Or In this seminar we... .

**CHAPTER 7**

**7. REFERENCES**

List of all the material that are used from various sources for making this project proposals:

*[1] Liu, Hong, Ashwin Kumar Tk, Johnson P. Thomas, and Xiaofei Hou. "Cleaning Framework for BigData: An Interactive Approach for Data Cleaning." In Big Data Computing Service and Applications (BigDataService), 2016 IEEE Second International Conference.*

*[2] Dr. Deepali Virmani, Preeti Arora, Ekta Sethi, Neha Sharma.*

*“Variegated Data Swabbing – An improved purged approach for data cleaning”. In 7th International Conference on Cloud Computing, Data Science & Engineering – Confluence*

*[3] Ling He, Zhongnan Zhang1, Yize Tan, Minghong Liao.*

*“An Efficient Data Cleaning Algorithm based on Attributes Selection. In 2011 6th International Conference on Computer Sciences and Convergence Information Technology (ICCIT)*

*[4] Virender Kumar, Cherry Khosla.*

*“Data Cleaning – A thorough analysis and survey on*

*Unstructured data” in 2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*

*[5] Berlian Al Kindhi, Muhammad Afif Hendrawan, Diana Purwitasari, Tri Arief Sardjono, Mauridhi Hery Purnomo.*

*“Distance-based Pattern Matching of DNA Sequences for Evaluating Primary Mutation” in 2017 IEEE 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*

*[6] S. Swapna, P. Niranjan, B. Srinivas, R. Swapna.*

*“Data Cleaning for Data Quality”. In 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*