****

**School of Computer Science & Engineering (SCOPE)**

**Project Topic: Efficient Clustering Techniques in Presence of Noise.**

**Course Name: Software Engineering**

**Course Code: CSE325**

**Slot : L43-44**

**Faculty :** Prof. Sathyaraj R.

**Submitted by:**

PURVA SINGH (14BCE0136)

KASHISH ARORA (14BCE0144)

SHIVIN SINHA (14BCE0130)

**ABSTRACT**

**AIM**: Implement efficient clustering technique for data set in presence of noise.

**ABSTRACT**: Mining Information and Knowledge patterns from large databases have been recognized by many researchers as key research topic in database systems, Knowledgebase systems, and statistics and in Information providing services. Clustering analysis method is one of the main analytical methods in data mining; the method of clustering algorithm will influence the clustering results directly. Clustering can be applied on database using various approaches based on distance, density, hierarchy and partition. The presence of Noise is a major problem in clustering. Noise is a data item that is not relevant to data mining. The Objective of the paper is present new algorithms for clustering techniques that handles the noise effectively. Our focus is to show the effect of noise on the performance of various types of clustering techniques and to study how noise affects the clustering process in terms of time and space. In the paper, implemented various clustering techniques such as K-Mediods and F C-Means. In this computed time complexity and space complexity of various clustering techniques for different number of clusters. These results are presented in various visual presentations like Line Chart, Bar Chart. Then we will conclude which algorithm is more efficient to deal noise. **STAKEHOLDERS:**

1) Administrator who manages the database.

2) Users of the system who wants to use the clustering data service (eg. Hospital staff, school management).

**LIST OF REQUIREMENTS:**

1) Users shall be able to select the clustering technique by selecting the options from the drop-down menu.

2) Users shall be able to view the types of clustering techniques available for the service.

3.)The output of the system will help the users to compare and contrast different clustering techniques.

**EFFICIENT CLUSTERING OF DATA IN THE PRESENCE OF NOISE**

**SOFTWARE REQUIREMENTS SPECIFICATION**

**DATE CREATED: 20/1/2017**

List of Contributors

| **Name** | **Initials** | **Project Title** | **Email** |
| --- | --- | --- | --- |
| Purva Singh | PS | Efficient Clustering Techniques in presence of noise | purva.singh2014@vit.ac.in |
| Kashish Arora | KA | Efficient Clustering Techniques in presence of noise | Kashish.arora2014@vit.ac.in |
| Shivin Sinha | SS | Efficient Clustering Techniques in the presence of noise | shivin.sinha2014@vit.ac.in |

**PREFACE:**

This document represents the Software Requirements Specification for the Efficient Clustering Techniques in presence of noise System. The document begins with an Introduction section that describes the purpose of the document and what is considered to be in the scope of this document as well as what is outside the scope of this document.

The next section is an Overall Description of the requirements and functions. This section includes the overall constraints that the project is working within as well as the assumptions made by the project as far as the defining the requirements is concerned. Lastly, the project dependencies are also listed in this section.

The Specific Requirements section comes next and is the most important section of this document. This section goes into detail about each specific requirement of the Efficient Clustering Techniques in presence of noise project. A description, use case with sequence of events, and any related requirements is given for each requirement.

Design Constraints and Standards Compliance are also considered in this section. Lastly, various System Attributes are discussed including Maintainability, Security, and Portability.

Table of Contents

[**1 Introduction..................................................................................................**1](#_Toc52901652)

[1.1 Purpose 1](#_Toc52901653)

[1.2 Scope 1](#_Toc52901654)

[1.3 Definitions, Acronyms, and Abbreviations 1](#_Toc52901655)

[**2 Overall Description** 2](#_Toc52901657)

[2.1 Product Perspective 2](#_Toc52901658)

[2.2 Product Functions 2](#_Toc52901659)

[2.3 Constraints 3](#_Toc52901660)

[2.4 User Characteristics 4](#_Toc52901661)

[**3 Problem Description** 12](#_Toc52901667)

**[4 Process Model](#_Toc52901668)**[……………………………………………………………….....................................12](#_Toc52901668)

[4.1 Why ? …………………………………………………………………………..........................................12](#_Toc52901668)

[4.2 Why not other ? ……………………………………………………………….....................................12](#_Toc52901668)

# 5 *Gantt Chart………………………………………………………………………....14*

5.1 Process based Gantt Chart……………………………………………………...................................14

## 5.2 Product based Gantt Chart……………………………………………………....14

# 6 Work Breakdown Structure…………………………………………………….......15

# 6.1 Activity Network………………………………………………………………….....15

**7 *List Of Stakeholders****…………………………………………………………….......................................15*

7.1 *Elicitation of Requirements……………………………………………………...................................15*

*7.2* *Functionalities…………………………………………………………………........................................…16*

**8** [**Performance Requirement….................................................................18**](#_Toc52901674)

8.1 Design Constraints………………………………………………………………......................................19

[8.2. Software System Attributes 19](#_Toc52901677)

[8.2.1 Reliability 19](#_Toc52901678)

[8.2.2 Security 20](#_Toc52901679)

[8.2.3 Maintainability 20](#_Toc52901680)

[8.2.4 Portability 20](#_Toc52901681)

List of Tables

1.3 [Table of Definitions, Acronyms, and Abbreviations 1](#_Toc52900262)

List of Figures

1. *Process Model………………………………………………………………..10*

## 5.1 Process based Grantt Chart…………………………………………………...11

## 5.2 Product based Grantt Chart……………………………………………………11

## 6 Work Breakdown Structure………………………………………………………..12

## 6.1Activity Network…………………………………………………………………....13

**1 INTRODUCTION:**

**1.1 PURPOSE:**

Research activity in clustering analysis initially focused on defining different types of algorithms to efficiently group large volumes of data in to clusters. The data to be clusterized is presented in database and in binary files.

To efficiently extract the data from disk and to divide into cluster from memory we have different cluster algorithms like

a) Hierarchical clustering algorithm

b) Partition algorithm

Some of these algorithms use distance as parameter and some other use density as parameter to cluster the data into groups. Later new techniques such as neural network based techniques, fuzzy techniques are developed to improve the clustering process efficiently. However noise is a problem in clustering process. The presence of noise reduces the performance of clustering process. So far no attempt is made to eliminate the noise. As a part of true integration we are presenting enhanced clustering techniques that eliminates problem of noise*.*

**1.2 SCOPE:**

“Efficient Clustering techniques in the presence of noise” is an application based on .NET framework which also includes user interaction. Our project is going to provide communication environment for users (testers, doctors). The GUI layer will be responsible for interaction with user and various calls to different graphical and visualization tools. This module also provides an interface to display performance results and comparison reports showing the performance of K-means, fuzzy c-means, K MEDIODS and enhanced C-means algorithms. The database server layer will use database server for maintaining transactional data items for use in K-MEDIODS clustering. As of now, we will be using ORACLE as data server.

In the end, the user will be able to compare the three different algorithms by seeing its time and computational complexity, by comparing the graph that displays clusters using fuzzy c means, k-means, K-MEDIODS and enhanced algorithms. In high level details, the system will use PHP for server side management, MYSQL DBMS to store and manipulate data and GUI to interact with users.

It is within the scope of this document to make user familiar to the the facilities available to them through this system developed in this project. The facilities include, generation of clusters from dataset by using K-Means, K-Mediods, enhanced fuzzy means and fuzzy C means, displaying time and space complexity of the mentioned algorithms, to visualize the generated line chart, bar chart and the table format displaying the time and space complexity analysis for different algorithms.

.

**1.3 Definitions, Acronyms, and Abbreviations**

|  |  |
| --- | --- |
| **TERMS** | **DEFINATION** |
| DBMS | Database Management System |
| RDBMS | Relational Database Management System |
| GUI | Graphical User Interface |
| KDD | Knowledge Discovery and Data Mining |
| DMQL | Data Mining Query Language |

**2 OVERALL DESCRIPTION:**

**2.1 PERSPECTIVE:**

The end user is provided with an interface to invoke with system and to execute queries based on clustering algorithms. The project also provides an interface for user to display the performance results and comparison reports showing performance of k-means, fuzzy c-means, k-mediods and enhanced fuzzy c-means algorithm. Finally the user will be able to see and compare different clustering algorithms based on line graphs, bar charts, comparison tables that compares and contrasts the time and space complexity of various algorithms as per data entered by user which is the main aim of this project

**2.2 FUNCTIONALITIES:**

The main functions of “Efficient Clustering techniques in the presence of noise” are:

* To design an interface for enhanced k-means clustering
* To design an interface for K-Medoids clustering
* To generate graph that displays clusters using enhanced k-means- algorithm
* To generate graph that displays clusters using KMedoids algorithm
* Computing time and space required to generate clusters.
* Generating graph showing clusters of KMedoids algorithm
* Generating graph showing clusters of CURE algorithm

The performance analysis layer presents the computational complexity and time complexity for generating clusters using k-means algorithm. The computational and time complexities are generated for different sizes of data.

The database server layer uses a database server for maintaining transactional data items used in k-mediods clustering.

The data base layer is a collection of database tables that maintains transactional data items. We can also use datasets for maintaining transactional data items. For implementing this system we are maintain student’s semester marks data.

**2.3 CONSTRAINTS:**

Our system does not provide security against database access. There is no login system in our system through which it can be ensured that only a specific member of the group can access the database and perform clustering algorithms over the data and modify it. So, this is an important constraint for our clustering system. Another important constraint is that the system requires a remote server which enables the system functionality and data storage. Because of this situation, when the server crashes the system will not be able to do its operations until the server become available to respond system requests. In addition to this, in case of hospitals for example, if they use this system for clustering the MRI scan of a particular patient, then all the patient’s details would be stored in the database. There is a possibility that the database can be hacked and the user information will no longer be private to the user. Clustering system has constraints in terms of regulatory, reliability, safety and security, but these constraints can be manageable.

**2.4 USER CHARACTERISTICS:**

User-1: Doctors: This software is similar to any other software and hence the user should be aware of using any software in general. The system under consideration has only the hospital doctors as users who can use brain images as input data for this system and can test on the image data that which clustering technique is best suited for clustering brain images with noise. The user does not need to have complete understanding of the complex functionalities and internal processing of the system.

User-2: Testers: This software is also well suited for testers or any end user who wants to check and analyze that which amongst the mentioned clustering techniques is best suited for the data that they have given as input to the system. For example, teachers can give student’s marks as input data to system and then analyze that which clustering method is suitable for data clustering*.*

**3 PROBLEM DESCRIPTION:**

Clustering analysis method is one of the main analytical methods in data mining; the method of clustering algorithms will influence the clustering results directly. Clustering can be applied on database using various approaches based on distance, density, hierarchy and partition. The presence of Noise is a major problem in clustering. Noise is a data item that is not relevant to data mining. The Objective of the paper is present new algorithms for clustering techniques that handles the noise effectively. Our system shows the effect of noise on the performance of various types of clustering techniques and to study how noise affects the clustering process in terms of time and space.

**4 PROCESS MODEL:**

The model used in our project is the waterfall model. As said in the problem description, we want the system to be user friendly so that everyone can use it smoothly. Waterfall model allows us this opportunity.****

**4.1** ***why the waterfall model?***

The waterfall model is used here because:

* The project needs to be easy to use and understand so that everyone can avail the service easily.
* The requirements of this service are already known to us and are not much likely to change as the number of clusters and users using this system are limited and there is a limit to the number of requests.
* Only when a request for service is submitted and the form is completely filled without leaving any mandatory fields empty, then only the user can see the results, bar charts and comparison tables. And only then the admin can have a check at the databases.

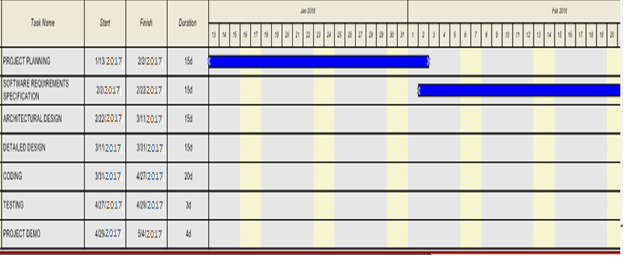
**4.2 why not other process model?**

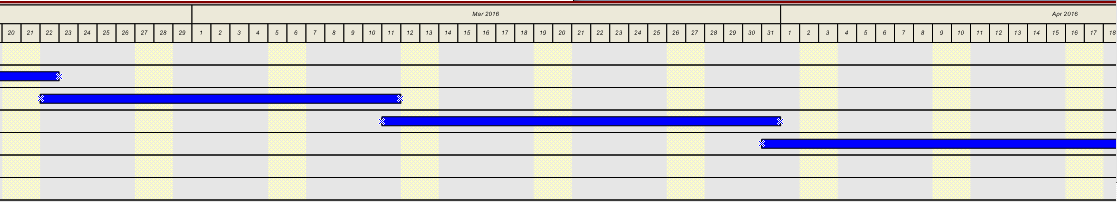
* The main reason behind choosing waterfall model is that the requirements and resources required for this project is very well known and is very less likely to change.
* An **incremental model** is not chosen as the requirements and resources available are known beforehand. Thus, there is no need for an incremental model where we get to know about the availability of resources with time.
* The **spiral system** is very time consuming as prototypes need to be created and tested constantly to know the requirements. In the case of our project this is futile as the requirements from the system are known already.
* **Evolutionary model** has a high skill requirement thus not used here.
* Project budget is not a factor as very less resources are required so **time box model** is not used.

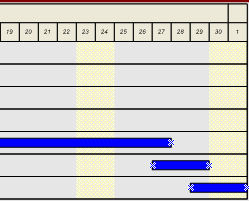
**5 GANTT CHART:**

**5.1 PROCESS BASED GANTT CHART:**

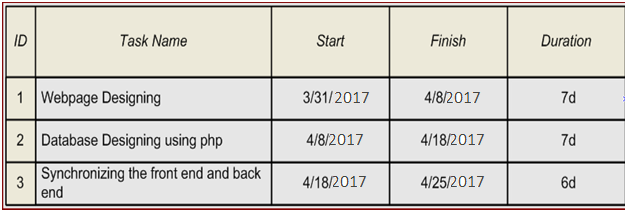
****

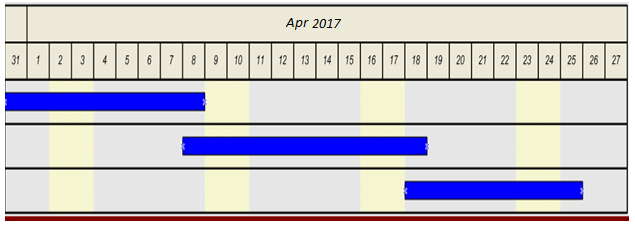
****

****

****

**5.2 PRODUCT BASED GANTT CHART:**

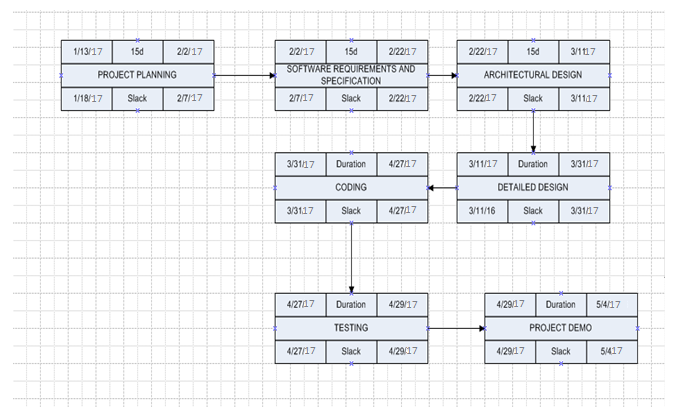
****

****

**6 WORK BREAKDOWN STRUCTURE:**

|  |  |  |
| --- | --- | --- |
| **TASK NAME** | **PREDECESSOR** | **DURATION** |
| PROJECT PLANNING |  | 15d |
| SOFTWARE REQUIREMENT SPECIFICATION | PROJECT PLANNING | 15d |
| ARCHITECTURAL DESIGN | SOFTWARE REQUIREMENT SPECIFICATION | 15d |
| DEATILED DESIGN | ARCHITECTURAL DESIGN | 15d |
| CODING | DEATILED DESIGN(OR) SOFTWARE REQUIREMENT SPECIFICATION | 20d |
| TESTING | CODING | 3d |
| PROJECT DEMO | TESTING | 4d |

**6.1 ACTIVITY NETWORK:**

****

**7 LIST OF STAKE HOLDERS:**

1) Management Staff (eg. Hospital management, school management) who manages the main system of clustering.

2) End-Users of the system (eg. Students, doctors) who want the clustering data service. The end-user interacting with the system generates binary association rule using binary association algorithm.

3) An analysis who analyzes existing business data in Data Base. The analysis also compares different algorithm performance to find out the efficient algorithm.

4) Authorized owners of the system (eg. HoD, dean, directors).

**7.1 Elicitation of Requirements:**

1) End-Users can report suggestions/complaints

2) Users shall be able to select the clustering technique by selecting the options from the drop-down menu.

3) Users shall be able to view the types of clustering techniques available for the service

4) End-user shall have a password protected login.

5) Administrator shall also have a password protected login.

6) Administrator shall be able to access the database to view the records.

7) User shall not be allowed to use the service without filling the mandatory fields in the form

9) User shall not be allowed to use the service without providing a proper path for dataset present in their computers or without providing the number of clusters he/she wants to form out of the data set.

10) The administrator shall be able to cancel a request and clear the database.

11) The user shall get the clusters formed in the result once he finishes entering the mandatory details.

12) User shall not be able to view the details of other users, the only permission to access all the databases is in the hands of the admin.

13)Though the user can see the time, space complexity and different comparison charts of various clustering techniques so as to compare their performance.

**7.2 Specific Requirements:**

1) Creating a page that asks the user whether he is a staff member or the administrator.

Functional Requirements:

* **PURPOSE**: To differentiate between the admin and other users. Since admin has special access to all the databases and other staff members do not have that permission, it is important to segregate the admin and the staff members.
* **INPUT**: The user will click on the check-box that will appear on his window, whether he belongs to the admin category or the staff category.
* **PROCESSING**: None
* **OUTPUT**: Will redirect the user to the login page of admin if he is an admin, or a staff login if he is a staff.

2) Creating the password protected form for the user to fill

* **PURPOSE**: To maintain privacy of data, this form will be password protected so that there is no breach of privacy among the users.
* **INPUT**: The user will enter the username and password that is assigned to him according to the category to which he belongs. This username and password is already predefined in the database and the user has to enter the exact same username and password to access the system.
* **PROCESSING**: The username and password entered by the user will be checked from the database and if they match completely, then the user will be granted access to the system.
* **OUTPUT**: The admin will see a list of database of the organization. The staff user will see a form that will begin the clustering process.

3) Creating the staff-form:

* **PURPOSE**: This page allows the user to input the details of the data set so as to apply various clustering algorithms to compare and contrast them on the basis of results.
* **INPUT**: The user will input the path to the data set in his computer (for eg. C:/desktop/dataset/set10) and then will enter the number of clusters he wants to form out of the data set.
* **PROCESSING**: This data is analyzed by the system and the result is shown to the user.
* **OUTPUT**: The user is directed to the results page where the user defined number of clusters are formed out of the data.

4) Creating the resultant graphs and table page:

* **PURPOSE**: This page is basically the second part of the result that user sees. The user will be able to compare the various clustering algorithm by analyzing the result.
* **INPUT**: This user selects the option of the type of chart he wants to see (for eg. Line or bar chart) or whether he wants to compare the results using table format.
* **PROCESSING**: The data filled in the form is analyzed.
* **OUTPUT**: The user will see a comparison table, the time taken by the algorithm for clustering for the noisy data. In this time taken for generating different clusters is high firstly and then it becomes uniform. Also the line and bar chart of time and space complexity of the different clustering algorithms is shown a result.

5) Creating the administrator

* **PURPOSE**: The admin can have a final look at the database of each and every user and can analyze it.
* **INPUT**: Details filled in by the user in the form given.
* **PROCESSING**: None
* **OUTPUT**: The admin has the final say as to which algorithm is best suited for their administration according to the type of data set that they use.

**8 PERFORMANCE REQUIREMENTS:**

The subsection specifies the following requirements associated with performance with which the Clustering System shall operate are:

**RESPONSE TIME:**

* The response time to load the system when the unit is turned on shall be less than 30 seconds.
* The response time to load the screen based on user’s choice should be less than 5 seconds.
* The response time to display the menu options on the display screen shall be less than 0.5 seconds.

**OPERATION:**

* The Navigation Bar shall provide menu options that will be easy to navigate by the user and should be user friendly.

* The application shall not crash or freeze the system, other than as the result of the operating system error. In such case, the user should again login to the application.

**SCALING**

* The system shall scale from 10 – 1000 users.
* The system shall support 1000 logged users.

# 8.1 DESIGN CONSTRAINTS

### 

### 8.1.1 SOFTWARE CONSTRAINTS

* The Clustering of Data in the presence of Noise system application shall be designed for future upgrade as per needed.
* The application shall meet the general standards of web applications.

### 8.1.2 HARDWARE CONSTRAINTS

* There is no hardware constraints identified at this point.

## 8.1.3 ACCEPTANCE CONSTRAINTS

To validate the system, the developers must complete the following:

* Demo the working system and any features upon request.
* Prove that all the significant functional requirements are met.
* Provide sufficient test cases to show that the system is complete and correct.

**8.2 SOFTWARE SYSTEM ATTRIBUTES**

**8.2.1 RELIABILITY**

The Clustering System never crash or freeze. The unit shall inform the user to re-set the system if major software mal-function has been encountered by the system.

**8.2.2 AVAILABILITY**

The application shall be available to any user who has internet connection.

**8.2.3 SECURITY**

The data entered by the user shall be kept secured and private and shall not be disclosed. The system shall be protected against hackers and otherwise malwares. The system shall make use security provided by the web server database. Firewall and network security are outside the scope of the document.

**8.2.4 MAINTAINABILITY**

The software should not require any additional maintenance. If any errors occur, the user should be able to logout and login again with his credentials. The system shall be flexible enough to add new modules and upgrade the existing modules.