

Title Page

Design & Implementation of Clustering techniques based Simulator

Project Report submitted in partial fulfillment of the requirement
for the degree of

Bachelor of Technology.

in

Computer Science & Engineering

under the Supervision of

Dr. Pardeep Kumar

By

Prateek Tiwari (091265)

to



Jaypee University of Information and Technology
Waknaghat, Solan – 173234, Himachal Pradesh

CERTIFICATE

This is to certify that project report entitled “ **Design & Implementation of Clustering techniques based Simulator**”, submitted by **Prateek Tiwari** in partial fulfillment for the award of degree of Bachelor of Technology in Computer Science & Engineering to Jaypee University of Information Technology, Waknaghat, Solan has been carried out under my supervision.

This work has not been submitted partially or fully to any other University or Institute for the award of this or any other degree or diploma.

Date: 12-05-2013

Supervisor's Name

Dr. Pardeep Kumar

Designation

Assistant Professor

ACKNOWLEDGEMENT

I would like to extend my sincere gratitude to our project mentor Dr. Paedeeep Kumar, Who provided us inspiring guidance and gave us an insight into the latest technical Development going in the vast field learning. I would also like him to thank for devoting considerable part his precious time to give us a better knowledge of the field data mining, without his support my project would not have been possible.

We would like to thak JUIT , Department of Computer Science & Engineering and all other members who provided me this golden opportunity and supported me to take another great step towards the completion final year project. I owe our heartiest thanks to Brig. (Retd.) S.P. Ghrrera (H.O.D CSE/I.T Department) who always inspired confidence in me to take initiative. He has always been motivating and encouragaing.

Date: **12-05-2013**

Name of the student

Prateek Tiwari

TABLE OF CONTENTS

S. No.	Topic	Page No.
1.	Introduction	1
1.1	Identification and significance of problem	1
1.2	Technical background	7
1.3	Data analysis task and techniques	9
2.	Data Mining	12
2.1	Need for data mining	13
2.2	Knowledge Discovery in Database	14
2.3	Data Mining and KDD	16
2.4	Types of Data Mining	17
3.	Clustering	18
3.1	Clustering Methods	19
3.2	Applications	21
3.3	Classification of Clustering Algorithm	21
4.	K Mean Method	23
4.1	History	23
4.2	Algorithms	24
4.3	Example	26
4.4	Features	32
4.5	Applications	35
5.	K Medoid Method	36
5.1	Proposed K Mediod Algorithms	37
5.2	Numerical Experiments	40
5.3	Difference Between K Mean and K Mediod Methods	45
6.	Code	47

7.	Application	52
7.1	Application on a Real Data Set	52
7.2	K Means Application	54

LIST OF FIGURES

S.No.	Title	Page No.
1.	Overview of the steps constituting the design approach process	9
2.	An overview of the knowledge discovery process	12
3.	An overview of the steps that compose the KDD Process	14
4.	The KDD Process	15
5.	An example of dendograms	21
6.	Clusters and Centroids after first pass	29
7.	Clusters and centroids after 2 nd pass	30
8.	A typical example of the k-means convergence to a local minimum	33
9.	K-mean example	34
10.	Artificial Data for Comparison	40
11.	(a) True cluster solution (b) Cluster result from K-mean (c) Cluster result from PAM and proposed method	43
12.	Time comparison of proposed method with PAM	44
13.	Overall performance versus cluster size (# of students) K=3	58

14. Overall performance versus cluster size (# of students) $K=4$	59
15. Overall performance versus cluster size (# of students) $K=5$	61

LIST OF TABLES

S.No.	Title	Page No.
1.	Finding_the nearest cluster center for each record. (First pass)	27
2.	The cluster_and centroids at the end of 2 nd pass	29
3.	The cluster and centroids at the end of 3 nd pass	30
4.	Mean and variance when generating objects	40
5.	Adjusted Rand indices by various clustering methods	41
6.	Comparison of K-means and K-mediods	45
7.	Records and Time	51
8.	Clusters and Time	52
9.	Performance Index	55
10.	For K=4	55
11.	For K=5	56

12.	Statistics of Data used	56
13.	For $K=3$	58

Abstract

Data mining and Knowledge discovery has several important application areas. Data mining and knowledge discovery have been topics considered at many AI, database and statistical conferences. Knowledge discovery generally refers to the process of identifying valid, novel and understandable patterns. Knowledge discovery from large databases, often called data mining, refers to the application of the discovery process on large databases or datasets. The discovery process can be broken into several steps, including: developing an understanding of the application domain; creating a target data set; data cleaning and preprocessing; finding useful features with which to represent the data; data mining to search for patterns of interest; and interpreting and consolidating discovered patterns. Data mining and knowledge discovery in databases have been attracting a significant amount of research, industry, and media attention of late. What is all the excitement about? This article provides an overview of this emerging field, clarifying how data mining and knowledge discovery in databases are related both to each other and to related fields, such as machine learning, statistics, and databases. The article mentions particular real-world applications, specific data-mining techniques, challenges involved in real-world applications of knowledge discovery, and current and future research directions in the field.

Data mining on large databases has been a major concern in research community, due to the difficulty of analyzing huge volumes of data using only traditional OLAP tools. This sort of process implies a lot of computational power, memory and disk I/O, which can only be provided by parallel computers. We present a discussion of how database technology can be integrated to data mining techniques. Finally, we also point out several advantages of addressing data consuming activities through a tight integration of a parallel database server and data mining techniques