A Project Report
on

Analysis of similarity of patterns using clustering techniques

Submitted for partial fulfillment of award of

BACHELOR OF TECHNOLOGY

Degree
In
Computer Science & Engineering

By

STUDENTS
Megha Pradhan (1622210084)
Bhavana Rai(1622210038)
Shilpi Singh(1622210156)
Karishma Kumari(1622210067)

Under the Guidance of
(Name of Guide)
Assistant Professor
Department of CSE

I.T.S ENGINEERING COLLEGE, GREATER NOIDA

APRIL 2020

CERTIFICATE

Certified that MEGHA PRADHAN, BHAWANA RAI, SHILPI SINGH, KARISHMA KUMARI has carried out the Project work presented in this project entitled **"ANALYSIS OF SIMILARITY OF PATTERNS USING CLUSTERING TECHNIQUES"** for the award of **Bachelor of Technology** from Dr. A.P.J. Abdul Kalam Technical University, Uttar Pradesh, Lucknow under my supervision. The Project embodies result of original work and studies carried out by Student himself and the contents of the Project do not form the basis for the award of any other degree to the candidate or to anybody else.

SANDEEP KUMAR
Assistant Professor
Department of CSE
Date:3/08/2020

Dr. Ashish Kumar
HOD
(Department of CSE)

ACKNOWLEDGEMENT

It gives us a great sense of pleasure to present the Report of the Project " Analysis of similarity of patterns using clustering techniques" undertaken during B.Tech final Year. First and foremost We wish to thank our Guide **Prof. Sandeep Kumar , Department of Computer Science and Engineering , I.T.S. Engineering College, Greater Noida** for his kind blessings to us . He allowed us the freedom to explore, while at the same time provided us with invaluable sight without which this Project would not have been possible.

We also do not like to miss the opportunity to acknowledge the contribution of all faculty members of the Department for their kind assistance and cooperation during the development of our project.

Megha Pradhan
(1622210084)

Bhavana Rai
(1622210038)

Shilpi Singh
(1622210156)

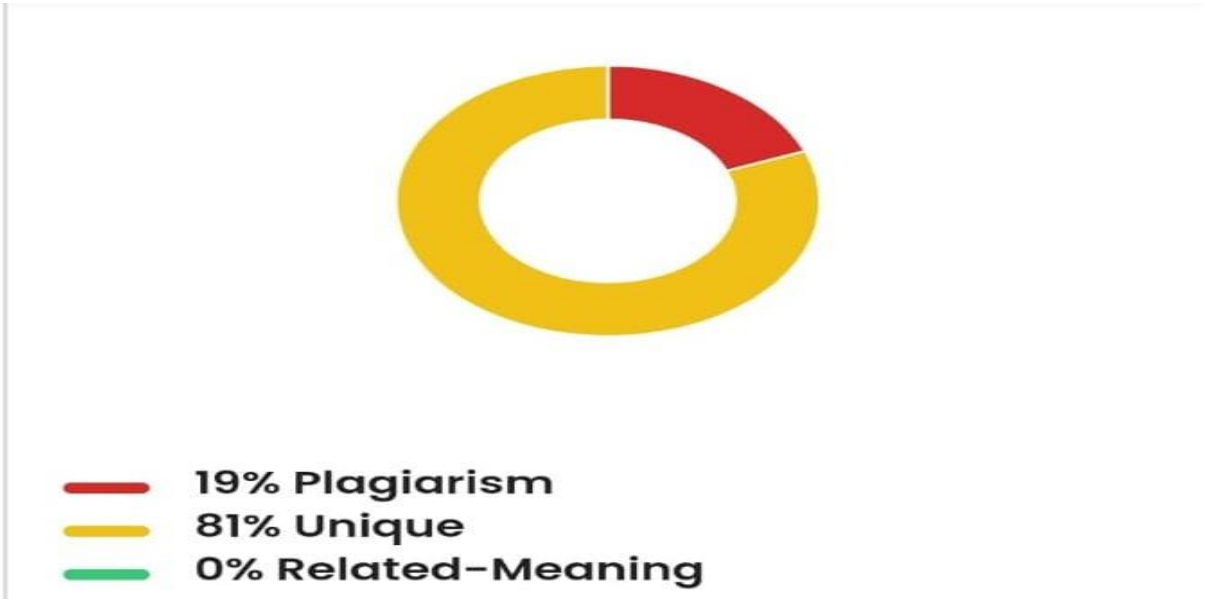
Karishma Kumari
(1622210067)

PLAGIARISM CERTIFICATE

This is hereby inform you that the project that we have done has the plagiarism certificate approved also. The result that we have got from the free online plagiarism checker is that there is 19% plagiarism , 81% unique and 0% related-meaning.

The website that we have used to check the plagiarism of our project named “ANALYSIS OF SIMILATITY IN PAATERNS USING CLUSTERING TECHNIQUES” is www.Duplichecker.com.

We have also attached the certificate of it below.



\

I.T.S ENGINEERING COLLEGE, GREATER NOIDA

CERTIFICATE OF PROJECT REPORT SUBMISSION FOR EVALUATION

Project Title: ANALYSIS OF SIMILARITY IN PATTERNS USING CLUSTERING TECHNIQUES

2. Project Preparation Guide was referred for preparing the Report ☐ YES ☐ NO
3. The contents of the Project Report have been organized based on the guidelines. ☐ YES ☐ NO
4. The Report has been prepared without resorting to plagiarism. ☐ YES ☐ NO
5. All sources used have been cited appropriately in Project Report ☐ YES ☐ NO
6. Submitted Two Hard bound copies along with one CD. ☐ YES ☐ NO

Megha Pradhan (1622210084)

Bhavana Rai (1622210038)

Karishma Kumari (1622210067)

Shilpi Singh (1622210156)

I.T.S ENGINEERING COLLEGE, GREATER NOIDA

ABSTRACT

The project relies on similarities in pattern. Some of the patterns from the nature is to be analysed like Qutub Minar, Indian Carpet, Barnsley leaf. In all the patterns, there are similarities in blocks of Qutub minar, patterns in leafs and symmetry view in the carpet. All the similarities are going to be analysed by clustering technique i.e. K- Means clustering technique.

Clustering is an unsupervised learning technique which aims at grouping a set of objects into clusters so that objects in the same clusters should be similar as possible, whereas objects in one cluster should be as dissimilar as possible from objects in other clusters.

Therefore, in this project we have considered the nature patterns and further to analyse the similarities using k-means clustering techniques.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

In the real world, there are so many patterns that have so many similarities. Some of the patterns from the nature is to be analyzed like Meenakshi Temple, Indian Carpet, Barnsley leaf. In all the patterns, there are similarities in blocks of temple, patterns in leafs and symmetry view in the carpet. All the similarities are going to be analyzed by clustering technique i.e. K- Means clustering technique.

Clustering is an automatic learning technique aimed at grouping a set of objects into subsets or clusters. Since there is not standard text classification criterion, it is very difficult for the people to use the massive text information sources effectively. Therefore, the management and analysis of text data become very important. Database management system gives access to data store but this was only a small part of what could be gained from the data. Analyzing the data by various techniques helped to gain further knowledge about the data explicitly stored to derive knowledge about the topic. This is where data mining or knowledge comes into existence.

1.1 DATA CLUSTERING

Data clustering is a popular data analysis task that involves the distribution of unannoted data (i.e, with np priory class information),in a manner into infinite sets of categories or cluster such that within a cluster are similar in some aspects and dissimilar from those in other clusters. Besides the term data clustering (or just clustering), there are a number of terms with similar meanings including cluster analysis automatic classification, numerical taxonomy and typological analysis. Clustering is defined as the problem of classifying in a collection of objects into a set of natural clusters without any prior knowledge. Data clustering attempts to discover and emphasis structural relationship between data vectors.

With the exponential growth of information and also a quickly growing number of text and hypertext document managed in organizational intranets represent the accumulated knowledge of organization that becomes more and more success in today's information society. Since there is not standard text classification criterion it is very difficult for people to use the massive text information source effectively. Therefore the management and analysis of text data become very important, nowadays such fields of text mining information filtering and information retrieving have brought great attention to both domestic and foreign expert. Document clustering aims to automatically group related documents into clusters, it is one of the most important tasks in machine learning and artificial intelligence and has received much attention in recent years.The main emphasis is to cluster with a high accuracy as possible. Document clustering has many important applications in the area of data mining and information retrieval. While doing the clustering analysis, we first partition the set of data into groups based on data similarity and then

assign the labels to the groups. The different algorithms are used for clustering the documents and to improve the quality to a great extent.

Clustering involves grouping of data into similar groups so that data in similar groups shares similar trends and patterns. Pattern within a valid cluster are more similar to each other than they are to a pattern belonging to a different cluster. There are different kinds of clusters such as compact, linear and circular. Clusters are formed based on the distance that is the points which are close are clustered in the same group which is called distance based clustering. A mathematical definition of a cluster is as follows: Let X be a set of data, that is X={ x1, x2,…, xn} A m-clustering of X is partitions into m parts (clusters) C1, C2,…., Cm, so that 1. None of the clusters is empty; Ci ≠Ø, 2. Every sample belongs to a cluster, 3. Every sample belongs to a single cluster; Ci ∩ j=Ø, i ≠ j.

Naturally it is assumed that vectors in a cluster Ci are in some way “more similar” to each other than to the vectors in other clusters.

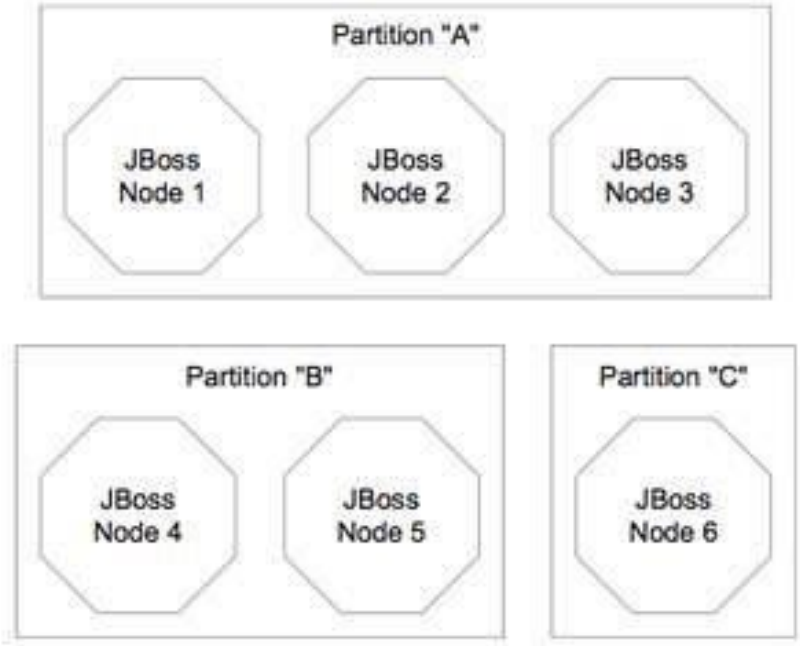


Fig.1.1 Pictorial Representation of Clustering

1.2 REQUIREMENTS

The main requirements that a clustering algorithm should satisfy are:

- Scalability
- Capable of dealing with different types of attributes
- Capable of discovering clusters
- Minimal requirements for domain knowledge to detrmine input parameters
- Ability to deal with noise and outliers
- High dimensionality
- Interpretability
- Usability

1.3 GOAL OF CLUSTERING

The goal of clustering is to decide the intrinsic grouping in a set of not labeled data. It can be shown that there is no absolute “best” criterion which would be independent of final aim of the clustering. Consequently it is the user and the real world application which must supply this criterion, in such a way that the results of the clustering will suit their needs.

CHAPTER 2

BACKGROUND AND RELATED WORK

2.1 LITERATURE REVIEW

Clustering objects into groups is usually based on a similarity metric between objects, with the goal that objects within the same group are very similar, and objects between different groups are less similar. In this review we focus on document clustering for web pages and tweet data. The application of text clustering can be both online or offline. Online applications are considered to be more efficient compared to offline applications in terms of cluster quality, however, they suffer from latency issues. Text clustering algorithms may be classified as flat clustering and hierarchical clustering. In the next two subsections we elaborate more details about these algorithms. Flat clustering algorithms Flat clustering explains how to create a flat set of clusters without any explicit structure that would relate clusters to each other. Flat clustering methods are conceptually simple, but they have a number of drawbacks. Most of the flat clustering algorithms, like k-means, require a pre-specified number of clusters as input and are non-deterministic. Hierarchical clustering algorithms Hierarchical clustering builds a cluster hierarchy, or in other words, a tree of clusters. Hierarchical clustering output is structured and more informative than flat clustering. Hierarchical clustering algorithms are further subdivided into two types:

(1) Agglomerative methods - a bottom-up cluster hierarchy generation by fusing objects into groups and groups into higher clusters.

(2) Divisive methods - a top-down cluster hierarchy generation by partitioning a single cluster encompassing all objects successively into finer clusters. Agglomerative techniques are more commonly used. Hierarchical clustering does not require knowing the pre-specified number of clusters. However this advantage came with the cost of the algorithm complexity. Hierarchical

The last stage of the clustering process is merging 4 base clusters to form the final clusters. K-Means is a generic clustering algorithm that can also be applied to clustering textual data. As opposed to Lingo and STC, bisecting k-means creates non-overlapping clusters.

Carrot2 is suited for clustering small to medium collections of documents. It may work for longer documents, but processing times will be too long for online search. The integration between Solr and Carrot2 is implemented as APIs. Learning about Solr-Carrot2 integration will help us in integrating our clustering techniques with Solar. Data Collection We evaluate clustering techniques on various tweet and web page collections. The collections include small data sets (< 500MB) and big data sets (> 1GB) and are related to various events of historical importance such as Ebola outbreak, Charlie Hebdo shooting incident, various incidents that took place on January, 25, Plane crash incident, Winter storm, Suicide bomb attack, Elections, Diabetes, tweets related to Egypt, Malaysia Airlines, Shooting, Storm, and Tunisia. The web pages for corresponding events are crawled using web links in each of the tweet collection. Web pages and tweets clustering In clustering of web pages, clustering approaches could be classified in two broad categories: term-based clustering and link-based clustering. Term-based clustering is based on common terms shared among documents. However, it does not adapt well to the web environment since it ignores the availability of hyperlinks between web pages. Link-based clustering could cluster web pages based on the information in the link. However, it suffers from the fact that pages without sufficient information in the links could not be clustered. It is natural to combine link and content information in the clustering algorithms to overcome the problems.

For tweets, a standard document clustering algorithms can be used. One interesting point in tweet clustering is the automatic detection of tweet topics, for example, the hash-tags that appear in tweets can be viewed as an approximate indicator of a tweet topic. Mahout clustering Mahout provides both in-memory and map-reduce versions of various clustering algorithms.

These algorithms are K-Means, Canopy, Fuzzy K-Means, and streaming k-mean, and Spectral Clustering. All these algorithms expect data in the form of vectors, so the first step is to convert the input data into this format, a process known as vactorisation. Essentially, clustering is the process of finding nearby points in n-dimensional space, where each vector represents a point in this space, and each element of a vector represents a dimension in this space. It is important to choose the right vector format for the clustering algorithm. For example, one should use the “Sequential Access Sparse Vector” for k-means. Other possibilities are the “Dense Vector” and the “Random Access Sparse Vector” formats. The input to a clustering algorithm is a sequence file containing key-value pairs of objects.

CHAPTER 3

K-MEANS

K-Means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed prior.. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more. Finally, this algorithm aims at minimizing an objective function know as squared error function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\left\| x_i - v_j \right\| \right)^2$$

where,

'||xi - vj||' is the Euclidean distance between xi and vj.

'ci' is the number of data points in ith cluster.

'c' is the number of cluster centers.

Algorithmic steps for k-means clustering

Let X = {x1,x2,x3,……..,xn} be the set of data points and V = {v1,v2,…….,vc} be the set of centers.

1) Randomly select 'c' cluster centers.

2) Calculate the distance between each data point and cluster centers.

3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..

4) Recalculate the new cluster center using:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

where, 'ci' represents the number of data points in ith cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.

6) If no data point was reassigned then stop, otherwise repeat from step 3).

3.1 K-MEANS EVALUATION

Our evaluation approach is iterative and we aim to produce clustering results that can be improved over time by optimizing feature vectors and clustering algorithms. To ensure best results our approach is to: research → identify improvements → implement (or find an equivalent open source implementation) → integrate and evaluate. The end goal will be to document our experience with various clustering algorithms and techniques for identifying feature vectors using different set of data which ensures quality clustering with optimal performance.

3.2 K-MEANS LABELING

Clusters that are obtained as a result of the clustering process need to be labeled appropriately in order to understand the purpose of each cluster and to evaluate the effectiveness of clustering. Cluster labeling is based on selecting words from each cluster and then use them to label the clusters. There are three ways of selecting words for cluster labeling. The first method (discriminative labeling) assumes the existence of a document hierarchy, either manually constructed and/or populated, or a hierarchy resulting from application of a hierarchical clustering algorithm.

Using chi square tests of independence at each node in the hierarchy starting from the root, we determine a set of words that are equally likely to occur in any of the children of a current node. Such words are general for all of the sub-trees of a current node, and are excluded form the nodes below. The second method (non-discriminative labeling) selects words which both occur frequently in a cluster or effectively discriminate the given cluster from the other clusters. While the third method, using titles for labeling clusters, is based on the titles 8 of the documents within each cluster to find the most representative words for the cluster. The work in briefly describe a

technique to label clusters based on how many times a feature is used in a cluster. By utilizing this information, and also drawing on knowledge of the code, short titles are manually selected for the obtained clusters. Although labeling is performed manually, they point out that the automatically developed feature summary of each cluster makes the labeling process much easier. In 2001 Tzerpos et al. emphasizes that a clustering algorithm should have certain features to make its output easier to comprehend. These features include bounded cluster cardinality, which ensures that any single cluster does not contain a very large number of entities, and effective cluster naming. They use a pattern based approach to recognizing familiar subsystem structures within large systems. The identified patterns are expected to occur in large systems with around 100 source files. The same pattern-based approach is used for cluster labeling. In 2003 Tonella et al. describe the use of keywords within web pages to cluster and label similar pages. Both single words and a contiguous sequence of two words i.e., bigrams are considered as representative keywords of a webpage. Clustering as well as cluster labeling are carried out on the basis of keywords within a webpage. Cluster labels are ranked according to inverse keyword frequency.

3.3 APPLICATION OF K-MEANS ALGORITHM

1. **Clustering Algorithm in Identifying Cancerous Data**

Clustering algorithm can be used in identifying the cancerous data set. Initially we take known samples of cancerous and non cancerous data set. Label both the samples data set. We then randomly mix both samples and apply different clustering algorithms into the mixed samples data set (this is known as learning phase of clustering algorithm) and accordingly check the result for how many data set we are getting the correct results (since this is known samples we already know the results beforehand) and hence we can calculate the percentage of correct results obtained. Now, for some arbitrary sample data set if we apply the same algorithm we can expect the result to be the same percentage correct as we got during the learning phase of the particular algorithm. On this basis we can search for the best suitable clustering algorithm for our data samples.

References

1) A Comparison of Fuzzy and Non-Fuzzy clustering Techniques in Cancer Diagnosis by X.Y. Wang and J.M. Garibaldi.
2) Probability Density Estimation from Optimally Condensed Data Samples by Mark Girolami and Chao He.

2. Clustering Algorithm in Search Engines

Clustering algorithm is the backbone behind the search engines. Search engines try to group similar objects in one cluster and the dissimilar objects far from each other. It provides result for the searched data according to the nearest similar object which are clustered around the data to be searched. Better the clustering algorithm used, better are the chances of getting the required result on the front page. Hence, the definition of similar object play a crucial role in getting the search results, better the definition of similar object better the result is.
Most of the brainstorming activities needs to be done for defining the criteria to be used for similar object.

References

1. Clustering Billions of Images with Large Scale Nearest Neighbor Search by Ting Liu, Charles Rosenberg and H.A. Rowley.

3. Clustering Algorithm in Academics

The ability to monitor the progress of students' academic performance has been the critical issue for the academic community of higher learning. Clustering algorithm can be used to monitor the students' academic performance. Based on the students' score they are grouped into different-different clusters (using k-means etc), where each clusters denoting the different level of performance. By knowing the number of students' in each cluster we can know the average performance of a class as a whole.

References

1) Application of k-means clustering algorithm for prediction of students' academic performance by O.J. Oyelade, O.O. Oladipupo and I.C. Obagbuwa.

4. Clustering Algorithm in Wireless Sensor Network's Based Application

Clustering Algorithm can be used effectively in Wireless Sensor Network's based application. One application where it can be used is in Landmine detection. Clustering algorithm plays the role of finding the Cluster heads(or cluster center) which collects all the data in its respective cluster.

References

1) Clustering of wireless sensor and actor networks based on sensor distribution and connectivity by Kemal Akkaya, Fatih Senel and Brian McLaughlan.
2) Wireless Sensor Network based Adaptive Landmine Detection Algorithm by Abhishek Saurabh and Azad Naik.

CHAPTER 4

PATTERNS TO BE ANALYSED

4.1 Qutub Minar

The **Qutb Minar**, is a victory tower" that forms part of the Qutb complex, a UNESCO World Heritage Site in the Mehrauli area of New Delhi, India.[3][4] The height of Qutub Minar is 72.5 meters, making it the tallest minaret in the world built of bricks.[5][6] The tower tapers, and has a 14.3 metres (47 feet) base diameter, reducing to 2.7 metres (9 feet) at the top of the peak.[7] It contains a spiral staircase of 379 steps.[8][1]

Its closest comparator is the 62-metre all-brick Minaret of Jam in Afghanistan, of c.1190, a decade or so before the probable start of the Delhi tower.[9] The surfaces of both are elaborately decorated with inscriptions and geometric patterns; in Delhi the shaft is fluted with "superb stalactite bracketing under the balconies" at the top of each stage.[10] In general minarets were slow to be used in India, and are often detached from the main mosque where they exist.

Fig. 4.1: QUTUB MINAR

4.2 INDIAN CARPET

A carpet is a textile floor covering typically consisting of an upper layer of pile attached to a backing. The pile was traditionally made from wool, but since the 20th century, synthetic fibers such as polypropylene, nylon or polyester are often used, as these fibers are less expensive than wool. The pile usually consists of twisted tufts which are typically heat-treated to maintain their structure. The term carpet is often used interchangeably with the term rug, although rugs are typically considered to be smaller than a room and not attached to the floor.

Carpets are used for a variety of purposes, including insulating a person's feet from a cold tile or concrete floor, making a room more comfortable as a place to sit on the floor (e.g., when playing with children or as a prayer rug), reducing sound from walking (particularly in apartment buildings) and adding decoration or colour to a room. Carpets can be made in any colour by using differently dyed fibers. Carpets can have many different types of patterns and motifs used to decorate the surface. In the 2000s, carpets are used in industrial and commercial establishments such as retail stores and hotels and in private homes. In the 2010s, a huge range of carpets and rugs are available at many price and quality levels, ranging from inexpensive, synthetic carpets that are mass-produced in factories and used in commercial buildings to costly hand-knotted wool rugs which are used in private homes of wealthy families.

Carpets can be produced on a loom quite similar to woven fabric, made using needle felts, knotted by hand (in oriental rugs), made with their pile injected into a backing material (called tufting), flat woven, made by hooking wool or cotton through the meshes of a sturdy fabric or embroidered. Carpet is commonly made in widths of 12 feet (3.7 m) and 15 feet (4.6 m) in the US, 4 m and 5 m in Europe. Since the 19th and 20th century, where necessary for wall-to-wall carpet, different widths of carpet can be seamed together with a seaming iron and seam tape (formerly it was sewn together) and fixed to a floor over a cushioned underlay (pad) using nails, tack strips (known in the UK as gripper rods), adhesives, or occasionally decorative metal stair rods. Wall-to-wall carpet is distinguished from rugs or mats, which are loose-laid floor coverings, as wall-to-wall carpet is fixed to the floor and covers a much larger area.

- Indian carpets are known for their high density of knotting.
- The designs of indian carpets are basically symmetrical either vertically or horizontally.

Fig. 4.2: Indian Carpet

4.3 BARNSLEY LEAF

Like the Sierpinski triangle, the Barnsley fern shows how graphically beautiful structures can be built from repetitive uses of mathematical formulas with computers. Barnsley's 1988 book Fractals Everywhere is based on the course which he taught for undergraduate and graduate students in the School of Mathematics, Georgia Institute of Technology, called Fractal Geometry. After publishing the book, a second course was developed, called Fractal Measure Theory. Barnsley's work has been a source of inspiration to graphic artists attempting to imitate nature with mathematical models.

The fern code developed by Barnsley is an example of an iterated function system (IFS) to create a fractal. This follows from the collage theorem. He has used fractals to model a diverse range of phenomena in science and technology, but most specifically plant structures.

Some facts about Barnsley fern are:

- The Barnsley fern is a fractal that is one of the basic examples ofself similar sets.
- It is a mathematically generated pattern that can be reproducible at any magnification or reduction.
- Fractal fern in four states of construction.
- Lifespan of fern dependson the species . Some types of ferns can live upto 100 years.

Fig. 4.3: Barnsley Leaf

CHAPTER 5

Design and Implementation

5.1 Data Set Description

Data sets are simple text documents from different domain. Given below are few documents, there are such many documents used for clustering. Following news group of data has been used for clustering:

Qutub Minar

=== Run information ===

Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10
Relation: output_file
Instances: 33925
Attributes: 4
R
G
B
Ignored:
img_name
Test mode: Classes to clusters evaluation on training data

=== Clustering model (full training set) ===

kMeans
======

Number of iterations: 33
Within cluster sum of squared errors: 2480.671382791246

Initial starting points (random):

Cluster 0: 164,119,96
Cluster 1: 129,77,55

Missing values globally replaced with mean/mode

Final cluster centroids:

		Cluster#	
Attribute	Full Data	0	1
	(33925.0)	(10535.0)	(23390.0)
R	173.3077	162.9799	177.9594
G	136.7641	188.0852	113.6488
B	124.1031	221.9416	80.036

Time taken to build model (full training data) : 0.58 seconds

=== Model and evaluation on training set ===

Clustered Instances

0 10535 (31%)
1 23390 (69%)

Class attribute: img_name
Classes to Clusters:

0	1	<-- assigned to cluster
2422	7658	image2.jpg
5987	10996	image3.jpg
2126	4736	image4.jpg

Cluster 0 <-- image3.jpg
Cluster 1 <-- image2.jpg

Incorrectly clustered instances : 20280.0 59.7789 %

Indian Carpet

=== Run information ===

Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10
Relation: indcarpet
Instances: 311252
Attributes: 4
R
G
B
Ignored:
img_name
Test mode: Classes to clusters evaluation on training data

=== Clustering model (full training set) ===

kMeans
======

Number of iterations: 16
Within cluster sum of squared errors: 10166.699477017733

Initial starting points (random):

Cluster 0: 112,102,75
Cluster 1: 160,136,98

Missing values globally replaced with mean/mode

Final cluster centroids:

	Cluster#		
Attribute	Full Data	0	1
	(311252.0)	(146326.0)	(164926.0)
R	170.8548	131.9716	205.3528
G	140.5352	93.4818	182.2821
B	107.6786	62.3463	147.8985

Time taken to build model (full training data) : 2.15 seconds

=== Model and evaluation on training set ===

Clustered Instances

0 146326 (47%)
1 164926 (53%)

Class attribute: img_name
Classes to Clusters:

0 1 <-- assigned to cluster
50474 53817 | image1.1.jpg
47142 55652 | image1.2.jpg
24774 28893 | image1.3.jpg
23936 26564 | image1.4.jpg

Cluster 0 <-- image1.1.jpg
Cluster 1 <-- image1.2.jpg

Incorrectly clustered instances : 205126.0 65.9035 %

Barnsley Leaf

=== Run information ===

Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10
Relation: leafcsv
Instances: 603160
Attributes: 4
R
G
B
Ignored:
img_name
Test mode: Classes to clusters evaluation on training data

=== Clustering model (full training set) ===

kMeans
======

Number of iterations: 7
Within cluster sum of squared errors: 23918.27257711945

Initial starting points (random):

Cluster 0: 103,136,57
Cluster 1: 149,181,105

Missing values globally replaced with mean/mode

Final cluster centroids:

		Cluster#	
Attribute	Full Data	0	1
	(603160.0)	(205039.0)	(398121.0)
R	188.1221	78.382	244.6402
G	216.322	153.5164	248.6679
B	178.7738	52.4729	243.8208

Time taken to build model (full training data) : 1.89 seconds

=== Model and evaluation on training set ===

Clustered Instances

0 205039 (34%)
1 398121 (66%)

Class attribute: img_name
Classes to Clusters:

0	1	<-- assigned to cluster
82989	167571	| image1.jpg
122050	230550	| image2.jpg

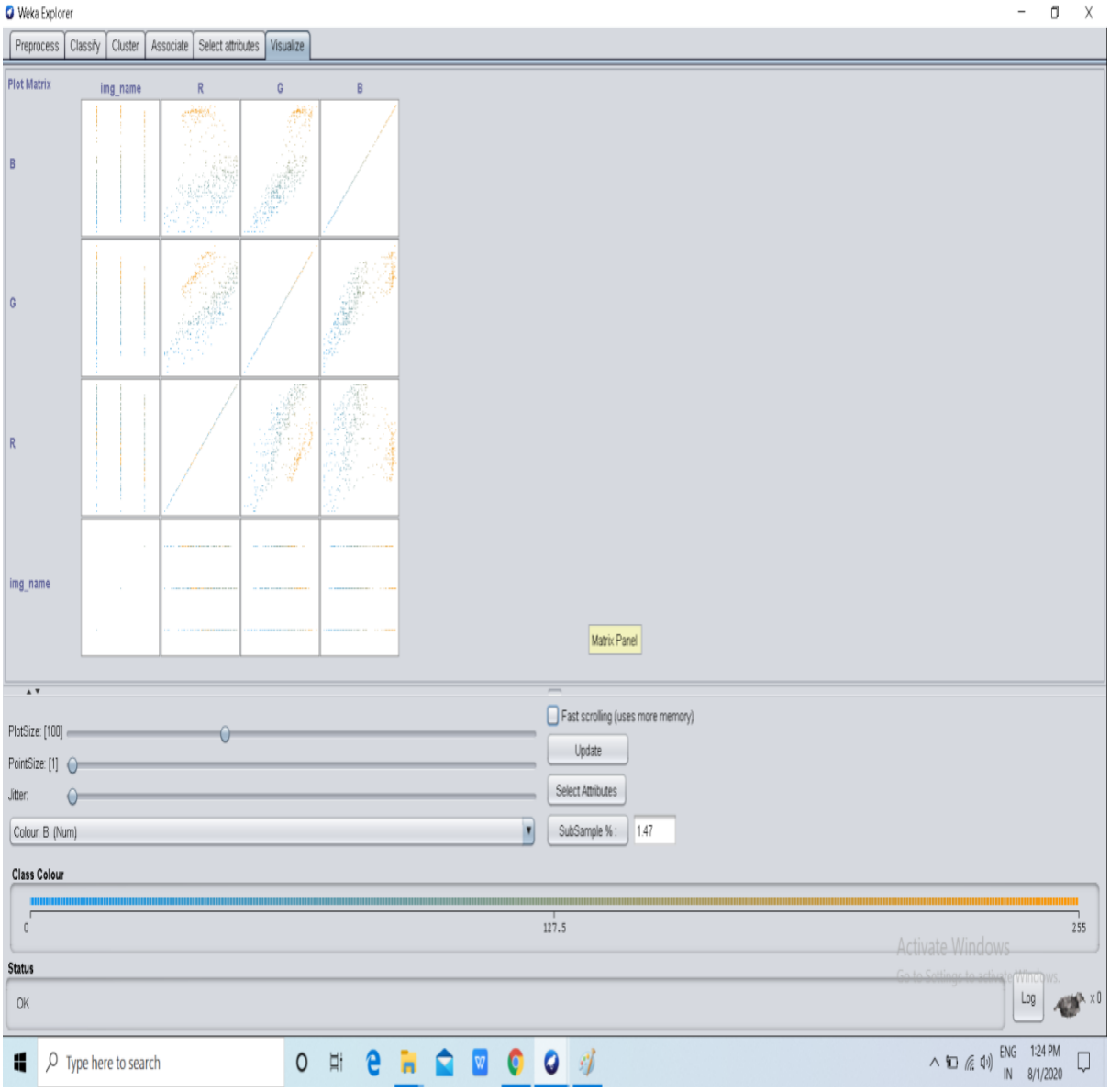
Cluster 0 <-- image1.jpg
Cluster 1 <-- image2.jpg

Incorrectly clustered instances : 289621.0 48.0173 %

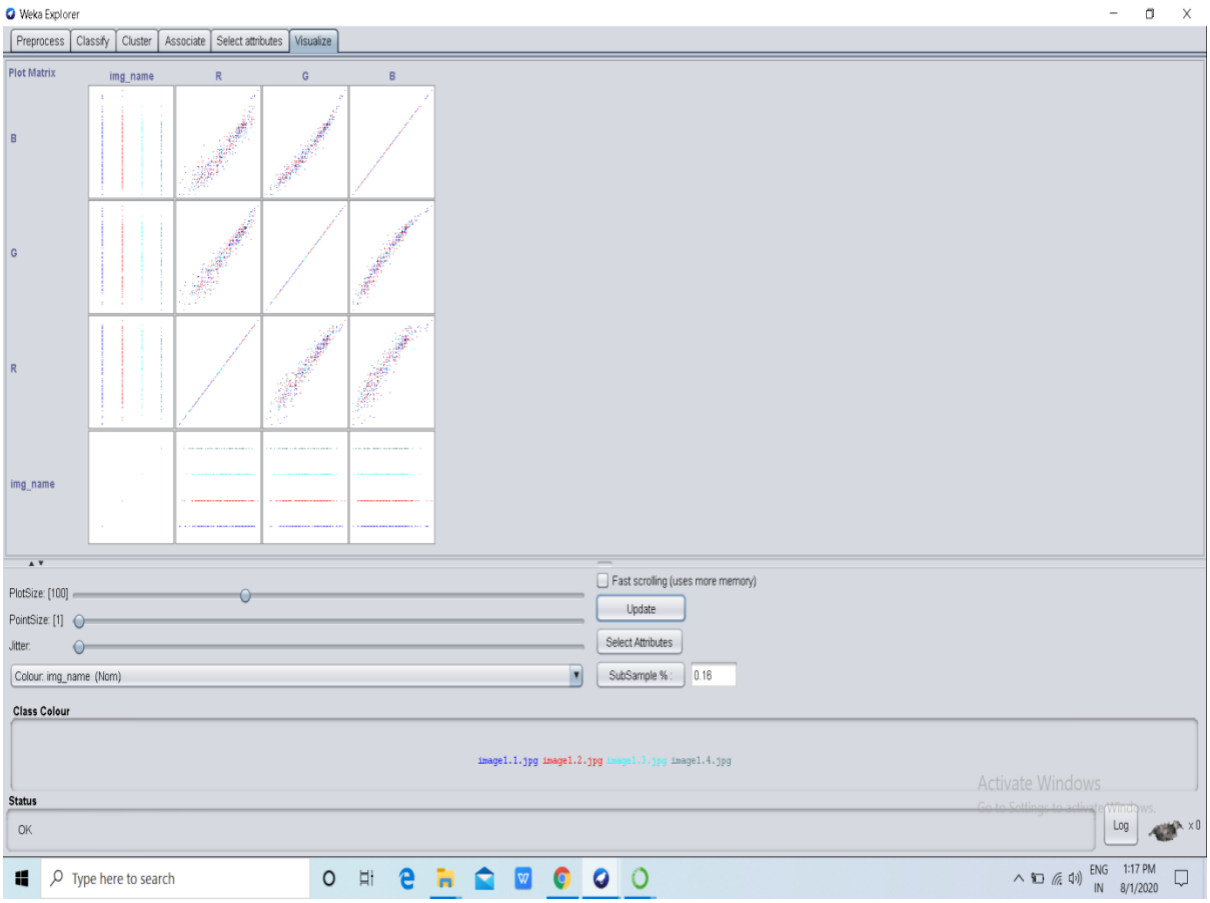
5.2 Screenshots

Given bellow all the screenshots of the clustering results:

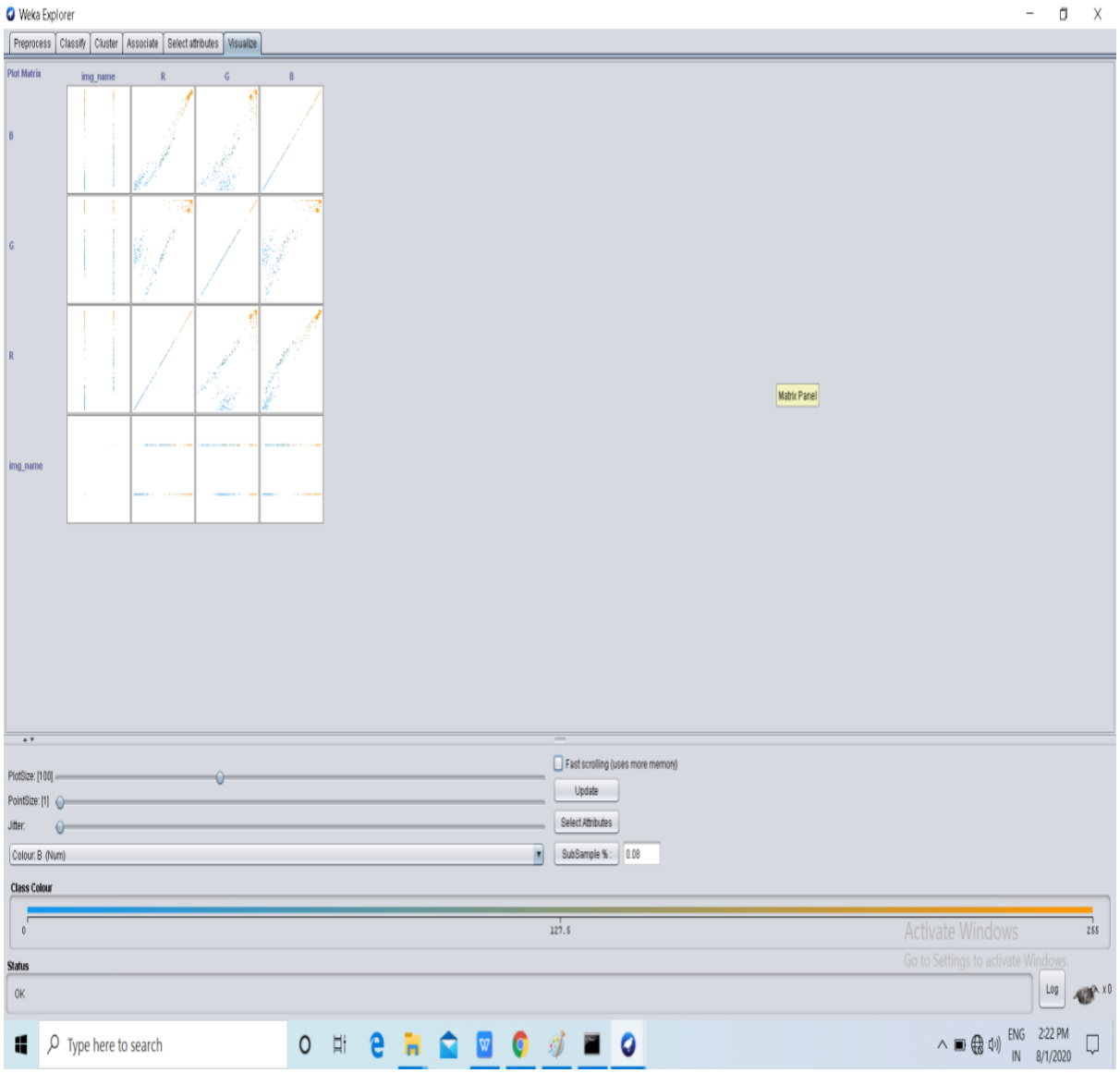
Qutub Minar


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Plot Matrix
img_name
R
G
B
Matrix Panel
Fast scrolling (uses more memory)
PlotSize: [100]
PointSize: [1]
Jitter:
Update
Select Attributes
Colour: B (Num)
SubSample %:
1.47
Class Colour
0
127.5
255
Status
OK
Log
x 0
Type here to search
ENG
IN
1:24 PM
8/1/2020

Indian Carpet



Barnsley Leaf



CHAPTER 6

Conclusion and Future Scope

In this we conclude that, is possible to have similarities in the nature patterns that we used in this project by using clustering techniques we can find the similarities. Thus we tried to implement the k-means algorithm which can be used to find the similarities in the clusters formed. The main contribution of this project is to:

Firstly we have find three patterns from the nature like qutub Minar, India carpet and
After that we have to check that if there is any similarity in it or not using k-means.

-> For that we have to convert the images of these patterns in to .csv file.

-> After doing that we have to give the input of the image that is in .csv extension into weka tool

-> By giving the input to the weka tool we will an output , for eg: Qutub minar we just have to take first section of the qutub minar as there is may be three or four sections in the qutub minar.

-> So now when we give the .csv value of the first section to the weka tool we will get the output of the first section and likewise we have to do it for all the remaining sections.

-> After getting all the ouputs of all the sections we will be able to find or we will be able to see that is there any similarity between the sections or not.

So, for future work, the mentioned clustering algorithms can be used for clustering and compare them to find out the best result.

BIBLIOGRAPHY

1) A Comparison of Fuzzy and Non-Fuzzy clustering Techniques in Cancer Diagnosis by X.Y. Wang and J.M. Garibaldi.
2) Probability Density Estimation from Optimally Condensed Data Samples by Mark Girolami and Chao He.
3) Clustering Billions of Images with Large Scale Nearest Neighbor Search by Ting Liu, Charles Rosenberg and H.A. Rowley.
4) Application of k-means clustering algorithm for prediction of students' academic performance by O.J. Oyelade, O.O. Oladipupo and I.C. Obagbuwa.
5) Clustering of wireless sensor and actor networks based on sensor distribution and connectivity by Kemal Akkaya, Fatih Senel and Brian McLaughlan.
6) Wireless Sensor Network based Adaptive Landmine Detection Algorithm by Abhishek Saurabh and Azad Naik.