Analysis and Clustering of PingER Network Data

Anwesha Mal
Department of CSE, ASET
Amity University
Noida, UP, India
anweshamal@gmail.com

Dr A.Sai Sabitha
Department of CSE, ASET
Amity University
Noida, UP, India
assabitha@amity.edu

Prof(Dr)Abhay Bansal
Department of CSE, ASET
Amity University
Noida, UP, India
abansal1@amity.edu

Prof(Dr)Bebo White
SLAC National Accelerator Laboratory
Stanford, CA, USA
bebo@slac.stanford.edu

Abstract— **The PingER project was started by SLAC National Accelerator Laboratory, Stanford, California for the purpose of monitoring end to end network data. For the last fifteen years pingER has generated an enormous amount of data that has been stored in CSV files in the form of Linked Open Data. However due to the difficulties faced in retrieving data efficiently, it has been proposed that all the data be put into the form of RDF triples. Interpreting and analysing such large volumes of data becomes a primary concern. By making using of clustering algorithms new and interesting patterns can be observed in the data sets. Outlier analysis can be performed giving insight to the exceptions occurring in the dataset and analysing the probable causes of such. Patterns could be observed based on the country to which the data belongs and comparisons can be drawn between the patterns between the different countries.**

Keywords— clustering, Network Monitoring, PingER, data analysis

I. Introduction

The pingER project was initially started in 1995 with the objective of aiding the High End Physics Community.. However in the recent years it has shifted its focus to measuring the divide in the digital world from the point of view of Internet Performance. It is an end to end internet monitoring tool. It was started by SLAC National Accelerator Laboratory, Stanford University, California. SLAC since then has collaborated with various other institutions to set up network monitoring sites all across the globe. When PingER originally started the Linked Open Data was stored in CSV files. However due to difficulty in accessing the data, this method of storing data was replaced with relational databases [2]. A system was devised to convert the flat files into relational database entries. This system faces issues with scalability and efficiency. To resolve this issue, it was suggested that data be stored in the format of RDF Triples which belongs to the World Wide Web Consortium standard. Though this is the format currently in use, it too faces some challenges. Another issue is the amount of data. PingER generates enormous amounts of data and analyzing this data becomes a challenging task. The objective is to find interesting and undiscovered patterns in the data by clustering data based on different parameters such as the country it belongs to [4]. The trends for various countries can be analyzed and compared and new conclusions can be drawn from them.

II. LITERATURE SURVEY

A. History of PingER

PingER is a project started by SLAC laboratories, Stanford. The basic objective is to monitor end to end pings in a network and study internet performance based on the speed of transfer of these pings. The project presently monitors around 700 sites from approximately 160 countries across the globe [1]. It was developed by the IEPM group at the Stanford Linear Accelerator Center (SLAC). The PingER data repository for network data is around a decade old and measurements from and to sites around the world are found in it. PingER data monitors approximately ninety nine percent of the internet population data. PingER had 20 monitoring sites all around the world in December 1999[9]. Eight of which were in the United States. Monitoring sites in Asia were located in Japan and China. Japan had two and China had one [12].Some of the details are listed in the table below:-

Table 1: Monitoring sites in USA

SLAC National Accelerator Laboratory	California
HEPNRC(HEP Network Resource Center)	Fermi National Laboratory , Illinois
Department of Energy	Washington DC
BNl	New York
Carnegie Mellon University	Pittsburgh
ARM (Atmospheric Radiation Measurement)	University of Maryland, College Park

In Europe there were seven other monitoring sites. The following table lists a few of them

Table 2: Monitoring Sites in Europe

Monitoring Site	Location
CERN	Geneva
DESY	Germany
INTN's National Centre for telematics and Information	

Ruderforth Library	

PingER comprises of several parts such as PingER validation, PingER operation, PingER analysis, PingER deployment, PingER Databases, PingER data and toolbox.

The monitored network data has been put to various uses such as rating the progress of a country based on its internet performance [11].For example SLAC managed to make conclusions about the quality of internet in Africa in the year 2013 by using the monitoring network data.

`

B. RDF Triples

RDF (Resource Description Framework) Triples are a format of data representation comprising of three parts which are a subject, a predicate and an object. The subject contains either a blank node or an URI reference. The predicate contains an RDF URI reference. The object is also a reference which could either be a blank node or a literal. It belongs to the family of World Wide Web consortium specifications. RDF triples are used because it makes data representation in a semantic web much simpler and organized. It puts all data into a common format which makes it simpler to integrate and combine the data. For this reason SLAC also attempts to put all the PingER data into RDF triple format.

C. Data Mining Techniques

Data Mining techniques are the methods with are used to analyse and find interesting patterns in the data. The data mining techniques in use are clustering, association, classification, and regression. Statistical methods are also implemented [12].

D. Clustering

Clustering basically refers to the grouping of data on the basis of certain common parameters and attributes. It is a statistical technique which is used. Clustering can be broadly classified into Connectivity based Clustering, Centroid Based Clustering, Distribution based clustering and Density based Clustering [10]. They include groups of data bundled together and small distances between each of the clusters. It can be thought of as a multi objective optimization problem. The most commonly used clustering algorithms are the K-means clustering algorithm, Hierarchical clustering (AGNES, DIANA) [13].

For the purpose of the work the K-means Clustering algorithm was implemented. It is a qualitative approach to clustering. For a given number of data points, the objective is to put them into k clusters such that each data point is associated with a cluster. The value of k is determined by the optimal value of the silhouette index. This index basically studies the distance of separation between the various clusters. The lesser the distance, the closer the clusters are to each other [14]. The clustering algorithm makes use of the squares of the

The K means algorithm has been summarized as follows:-

1) The Group of objects are partitioned into K clusters.
2) The seed point is calculated which is the mean point of the cluster.
3) Assignment of each individual object to the closest seed point.
4) Step 2 is repeated until there is no further movement in points

$$\sum_{m=1}^{k}\sum_{t_{mi}\in Km}(C_m - t_{mi})^2 \quad \text{———} \quad \textbf{eq. 1}$$

III. METHODOLOGY

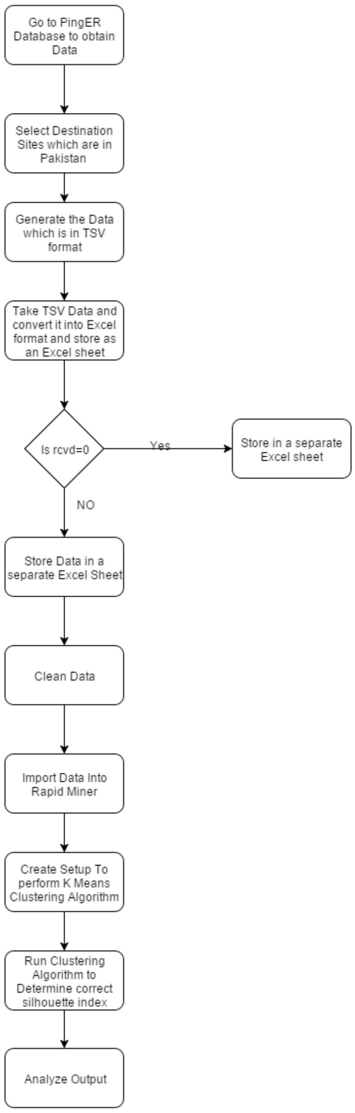
Go to PingER Database to obtain Data
Select Destination Sites which are in Pakistan
Generate the Data which is in TSV format
Take TSV Data and convert it into Excel format and store as an Excel sheet
Is rcvd=0
Yes
Store in a separate Excel sheet
NO
Store Data in a separate Excel Sheet
Clean Data
Import Data Into Rapid Miner
Create Setup To perform K Means Clustering Algorithm
Run Clustering Algorithm to Determine correct silhouette index
Analyze Output

Fig 1: Flowchart of process

1) Obtaining Data:-The required data was obtained from the official SLAC Laboratory. All sites having the.pk extension were considered and selected. The .pk extension indicates that the destination site is in Pakistan.

2) Conversion of Format: - The data from these respective sites were then then generated. The output obtained was in .tsv format (tab separated value). This was then converted into .xlsx format.

3) Data Cleaning:- The entries having the value of received as 0 were removed from the table.

4) Determining Silhouette Index:- The K means clustering algorithm was applied to the dataset for varying values of k. The silhouette index was obtained corresponding to every k value. The optimal was then chosen.

5) Applying K means Clustering and analysing results: - K means clustering algorithm was applied with the optimal K value and the results were analysed.

A. *Experimental Setup*

The data was collected from official SLAC website. All the sites containing the .pk extension was taken were considered and collected for analysis. The .pk extension indicates that the destination site is located in Pakistan. The data generated was in .tsv format (tab separated value) and was further converted into .xlsx format and then it is imported onto an excel sheet. The data obtained comprised of 9106 tuples. The various parameters in the data collected contained the following Attributes:-

Table 3:- Attributes of Dataset

source_host_name,	Name of source website
source_host_addr	IP address of source website
destination_host_name	Name od Destination Website
destination_host_addr	IP address of Destination Web Site
size	The size of the packets are fixed to be either 100 or 1000 bytes.
unix_epoch_time	The epoch time in the Greenwich Mean Time at which the measurement was made. It is used to calculate the absolute time in a unix system by eliminating the issue of time zones.
sent	the number of ping packets sent
rcvd	the number of ping packets received
min	the minimum of the ping packet Round Trip Delay Time Received
avg	the average of the ping packet Round Trip Delay Time Received
max	the maximum of the ping packet Round Trip Delay Time Received
seq_rcv(i=1,rcvd=10)	Refers to the sequence numbers of the packets received.
rtt_rcv(i=1,rcvd=10)	The RTT time of every packet received

B. *Data Cleaning*

1) The entries having the value of attribute received (rcvd) as zero were removed from the table and put into a separate table. It was observed that those entries that have an rcvd value of zero did not give any output.
2) All entries with received not equal to zero was considered for analysis. It was found that there were 4027 entries.
3) The attributes having constant value were then removed from the table. The attributes removed were the source_host_name, source_host_addr, size, unix_epoch_time, source_host_addr, source_host_addr. seq_rcv(i=1,rcvd=10), rtt_rcv(i=1,rcvd=10). Rtt_rcv was removed because the mix, max and average of all the packets were taken. Therefore it was unnecessary to take their values independently as well.

4) The attributes considered for analysis are the destination_host_name, destination_host_addr, sent, rcvd, min, avg, and max.

C. *Model Construction*

Rapid Miner Studio is used to perform cluster Analysis on the cleaned data set. Rapid Miner Studio is an open source tool which was used to perform basic K means clustering on the data.

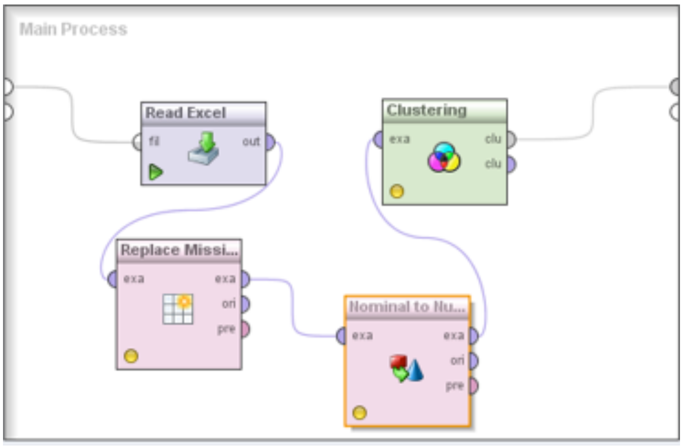

Fig 2:- Experimental Setup used in Rapid Miner

The model used in Rapid Miner to cluster the data is shown in figure 2.The excel sheet was imported and stored in a local repository. It was then fed it to a Nominal to Numerical Filter. This was done as the data set comprises of both alphanumeric and numeric entries and the clustering algorithm is unable to operate on such datasets. The Nominal to numerical filter resolves the problem of having different data types. The next step involved passing the data to a Replace Missing Value Filter as some of the data fields obtained were empty and the K means clustering algorithm would not work in such cases. After the data was prepared, K means clustering algorithm was applied to the dataset.

D. Optimal Value of K

To determine the optimal value of k, the clustering algorithm was run for various different values of k and the silhouette index was noted. The value of k was selected for the point at which the silhouette index was maximum.

Silhouette index is a measure which is used to determine the most appropriate value of k in K means analysis. It basically determines how well an item lies in a cluster. Silhouette values usually have a range from -1 to +1. They are also used to determine the distance between the various different clusters. The silhouette index is basically calculated as the mean of the intra cluster distance (a1) and the mean of the distance between the nearest clusters (b1). Therefore the silhouette index can be defined as the difference between b1 and a1 divided by max (a1,b1).

Silhouette Index = (b1-a1) / max (a1,b1)……eq.2

Table 4: Values of Silhouette Index

K=3	K=4	K=5	K=6	K=10
Average Silhouette Index is 0.631	Average Silhouette Index is 0.650	Average Silhouette Index is 0.689	Average Silhouette Index is 0.585	Average Silhouette Index is 0.586

From the above table it is clear that the optimal silhouette index is when k has a value of 5. The results are then obtained and then analyzed.

IV. Results And Analysis

K means clustering algorithm was applied to the data set comprising of 4027 entries with a k value of 5 as it gave the optimal silhouette index. The clusters obtained were as follows:-

Table 5:- Clusters and corresponding items

Cluster Number	Number of Items
Cluster 0	1
Cluster 1	124
Cluster 2	2
Cluster 3	906
Cluster 4	2993

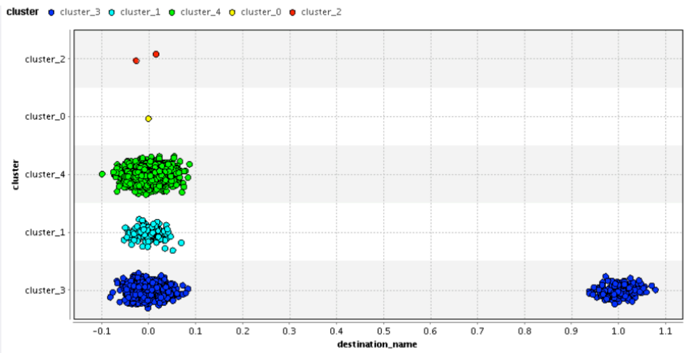

Fig 3: Plot of Clusters vs Destination name

The figure 3 shows the mapping of the various clusters with respect to the destination names. From the results it is evident that cluster 0 and cluster 2 are the outliers in the Analysis. Cluster 0 contains one entry. The corresponding entry has the destination name maggie2.seecs.edu.pk, its IP address is 115.186.131.82. 12 packets were sent to it and 10 were received in return. The minimum, average and maximum Round Trip Delay Time for the entry in Cluster 1 is 292.542, 832.618 and 3491.758. The value for maximum is unnaturally large here.

Cluster two contains two entries which can also be considered as outliers. Both the entries have the destination name maggie2.seecs.edu.pk and the IP address is 115.186.131.82. For both the entries the number of sent and received packets are 10. For the first entry the minimum is 412.42, the average is 543.414 and the maximum is 706.218. For the second entry the minimum is 724.236, the average 869.257 and the maximum is 1012.019.

It has been observed that all the outliers have seen generate by the maggie2.seecs.edu.pk.

For the next step of the analysis, the minimum, average and maximum were mapped against the clusters and plotted.

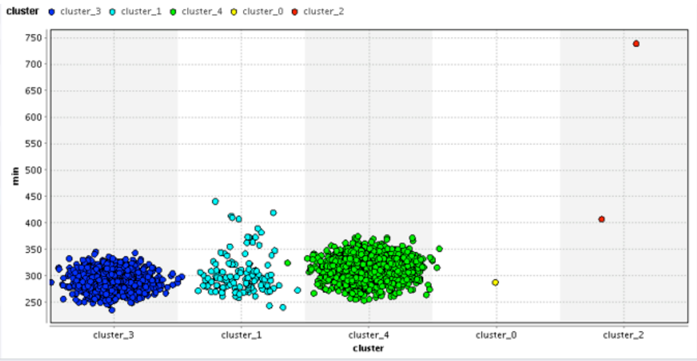

Fig 4: Plot of Min vs Cluster

Figure 4 shows the plot of the minimum against the clusters. It can be observed that the value of minimum for cluster 0 is in the same range as that of the other clusters which is approximately between 250 and 350. For cluster 2 it can be observed that one entry is only slightly above the average value whereas one point deviates greatly having a 724.236. Cluster 3 and Cluster 4 have closely clustered values with only a few points showing little deviation. On the other hand, the values in cluster 1 show quite a lot of deviation and the data values are not as closely clustered together as in cluster 3 and 4. The range of values in cluster 1 varies from approximately 240 to 450. One point can be seen to lie very close to cluster 4. The next analysis was performed by plotting the maximum values against the cluster.

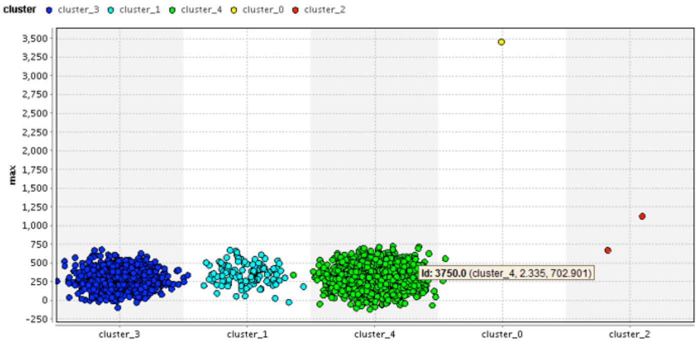

Fig 5: Plot of Max vs Cluster

Figure 5 shows the results obtained when the maximum was plotter against the cluster. Cluster 0 shows maximum deviation whereas while cluster 2 shows deviation compared to the other clusters, it is still less than that seen in cluster 2. Cluster 2 has a value which is approximately 3500. Cluster 1,3 and 4 are densely clustered showing little outliers. Though cluster 1 shows more compared to the other two. Here two there are some values in cluster 1 which lie very close to cluster 2. The variation in values for cluster 1 for the plot of maximum against destination is much less than the deviation observed when the minimum was plotted against the destination.

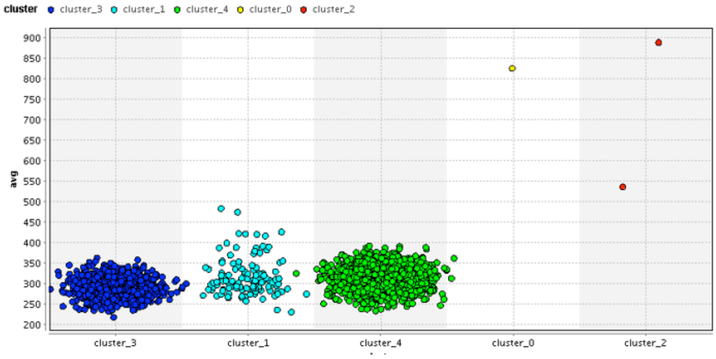

Fig 6: Plot of Avg vs Cluster

Lastly the average values were plotted against the clusters as shown in Figure 6. The above diagram shows the results of when the minimum is plotted against the clusters. Cluster 0 and Cluster 2 show great deviation in value of average. None of the points fall into the average range of values. Cluster 1 also shows quite a lot of deviation with approximately 15 data values varying from the average values of the other data values. The values in cluster 4 are quite densely grouped together with very little deviation. Cluster 1 too has points which all more or less lie in the same range with only very few points showing any significant deviation indicating that all the values are very close together.

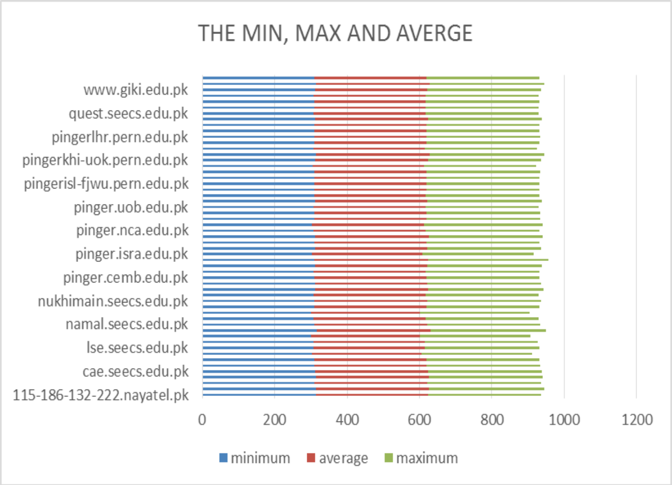

Fig 7: The min, max and average of all monitored sites

Fig 7 depicts all the destination monitoring sites which were all being monitored and shows the comparison between the minimum, maximum and average of the various sites being studied.

V. Future Work

The data collected for this paper can be further analyzed. Various other clustering algorithms can be implemented and the results obtained can be compared. The different variety of results may also be observed by changing the number of clusters in the K means clustering algorithm[8]. The work can be taken even further by analyzing data sets of different countries and studying the various trends observed in them. Trends between different countries can also be studied be studied and compared.

REFERENCES

[1] McCallum, Andrew, Kamal Nigam, and Lyle H. Ungar. "Efficient clustering of high-dimensional data sets with application to reference matching."Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2000.

[2] Anderson, Tessa K. "Kernel density estimation and K-means clustering to profile road accident hotspots." *Accident Analysis & Prevention* 41.3 (2009): 359-364.

[3] Wagstaff, Kiri, et al. "Constrained k-means clustering with background knowledge." *ICML*. Vol. 1. 2001.

[4] www1.slac.stanford.edu.html

[5] McCallum, Andrew, Kamal Nigam, and Lyle H. Ungar. "Efficient clustering of high-dimensional data sets with application to reference matching."*Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2000.

[6] Peng, Fuchun, and Andrew McCallum. "Information extraction from research papers using conditional random fields." *Information processing & management*42.4 (2006): 963-979.

[7] M Hoang, Ha, and Bostjan Antoncic. "Network-based research in entrepreneurship: A critical review." *Journal of business venturing* 18.2 (2003): 165-187.

[8] Eagle, Nathan, Alex Sandy Pentland, and David Lazer. "Inferring friendship network structure by using mobile phone data." *Proceedings of the National Academy of Sciences* 106.36 (2009): 15274-15278.

[9] Basu, Sugato, Ian Davidson, and Kiri Wagstaff, eds. *Constrained clustering: Advances in algorithms, theory, and applications*. CRC Press, 2008.

[10] Brieger, Ronald L. "Career attributes and network structure: A blockmodel study of a biomedical research specialty." *American sociological review* (1976): 117-135.

[11] Mantel, Nathan. "The detection of disease clustering and a generalized regression approach." *Cancer research* 27.2 Part 1 (1967): 209-220.

[12] http://www-iepm.slac.stanford.edu/pinger/#

[13] Ngai, Eric WT, Li Xiu, and Dorothy CK Chau. "Application of data mining techniques in customer relationship management: A literature review and classification." *Expert systems with applications* 36.2 (2009): 2592-2602.

[14] sNgai, Eric WT, Li Xiu, and Dorothy CK Chau. "Application of data mining techniques in customer relationship management: A literature review and classification." *Expert systems with applications* 36.2 (2009): 2592-2602.

[15] Ngai, Eric WT, Li Xiu, and Dorothy CK Chau. "Application of data mining techniques in customer relationship management: A literature review and classification." *Expert systems with applications* 36.2 (2009): 2592-2602.