Load Estimation of Social Networking Web Sites Using Clustering Technique

Deepti Bhagwani and Setu Kumar Chaturvedi
Department of Computer Science Engineering, Technocrat Institute of Technology, Bhopal, M.P., India
Email: {deepti.mca07, setu16}@gmail.com

Kapil Keswani
Department of Electronics & Communication Engineering, IPS College of Technology & Management, Gwalior, M.P., India
Email: profkeswani@gmail.com

Abstract—Facebook, Twitter and LinkedIn are the most popular online social networking sites on the Internet. These sites are a powerful mode of sharing, organizing and finding content and contacts. Usage of these sites is increasing so as to provide an opportunity to study the characteristics of online social networking sites at large scale. In this paper work, an attempt has been made to estimate the server load of social networking sites in order to maintain the servers efficiently. In this order, we have gathered the data for three popular social networking sites: Facebook, Twitter and LinkedIn from Internet Libraries. Datasets contain data of 600 cities across the world in terms of Number of users and response time respectively. Further, we have applied Dimension Reduction Algorithm to reduce the datasets for the purpose to attain the meaningful data. Thereafter, we have applied two clustering techniques (K-Means and Agglomerative hierarchical clustering) on these datasets to estimate the load of social networking sites. Results confirm that the clusters which arise from both the techniques contain various number of objects which specify that all the objects (i.e. cities) comes under that particular cluster cover same load to some extent that validate the hypothetical claims and exhibit the effectiveness of our algorithms.

Index Terms—online social networks, reduction algorithm, XLStat, K-mean clustering, agglomerative hierarchical clustering

I. INTRODUCTION

A. Online Social Networking

Online social networks are becoming a fast growing point of the Internet. As individuals continuously communicate with each other both in business as well as in personal contacts, the ability for the Internet to deliver this networking capability is becoming stronger and stronger. There are a number of resources available to anyone interested in becoming part of the online social networking community of the Internet. The term “social network” can be defined as a structure of social entities connected to other social entities through various types of relations [1]. We call these entities as “users”, for our specific purposes.

A social network is a set of people or organizations or other social entities connected by set of social relationships such as friendship, co-working or information exchange. Various algorithms and methods have been analyzed like Estimation Algorithms [2], Random Walk [3], Degree of Distribution [4], Link Prediction Technique [5], & Bayesian MCMC Method [6]. Table I shows the list of online social networking sites.

<table>
<thead>
<tr>
<th>OSNs</th>
<th>Number of Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flickr</td>
<td>1.8 million</td>
</tr>
<tr>
<td>Facebook</td>
<td>350 million</td>
</tr>
<tr>
<td>Orkut</td>
<td>100 million</td>
</tr>
<tr>
<td>Twitter</td>
<td>300 million</td>
</tr>
<tr>
<td>LinkedIn</td>
<td>50 million</td>
</tr>
<tr>
<td>YouTube</td>
<td>1.1 million</td>
</tr>
<tr>
<td>LiveJournal</td>
<td>5.2 million</td>
</tr>
<tr>
<td>Cyworld</td>
<td>48 million</td>
</tr>
<tr>
<td>MySpace</td>
<td>190 million</td>
</tr>
</tbody>
</table>

Figure 1. Data clustering

B. Clustering

Clustering is the unsupervised classification of patterns (data items, observations, or feature vectors) into clusters. A cluster can be defined as “A cluster is a set of entities which are similar and entities from different clusters are not similar.” A cluster is “an aggregation of points in the test space such that the distance between any two points in the cluster is less than the distance between any point within the cluster and any point outside the cluster.” Clearly, a cluster in above mentioned definitions is described in terms of internal homogeneity and external separation, i.e., data objects in the same cluster should be...
similar to each other as shown below in Fig. 1, although data objects in different clusters should be different from one another. [7]

1) **Clustering procedure**

As shown below in Fig. 2 the basic process of cluster analysis consists of four steps. These steps are closely related to each other and determine the derived clusters [8], [9].

- Feature selection or extraction, Feature selection chooses distinctive features from a set of candidates, while feature extraction uses some transformations to generate useful and novel features from the original ones.
- Clustering algorithm selection or design, this step usually consists of determining an appropriate proximity measure and constructing a criterion function.
- Cluster validation, given a data set, each clustering algorithm can always produce a partition whether or not there exists a particular structure in the data.
- Result interpretation, the ultimate goal of clustering is to provide users with meaningful insights from the original data so that they can develop a clear understanding of the data and therefore effectively solve the problems encountered.

![Figure 2. Steps of clustering procedure](image)

2) **Categorization of clustering algorithms**

Clustering algorithms differ among themselves in their ability to handle different types of attributes, numerical and categorical. Clustering can be performed both on numerical data and categorical data. To cluster numerical data, the inherent geometric properties can be used that define the distance between the points. But for clustering the categorical data, a criterion does not exist, on which distance functions are not naturally defined [10]. Clustering algorithms for numerical data and categorical data. To cluster numerical data, the inherent geometric properties can be used that define the distance between the points. But for clustering the categorical data, a criterion does not exist, on which distance functions are not naturally defined [10]. Clustering algorithms for numerical data and categorical data. To cluster numerical data, the inherent geometric properties can be used that define the distance between the points. But for clustering the categorical data, a criterion does not exist, on which distance functions are not naturally defined [10]. Clustering algorithms for numerical data and categorical.

3) **Hierarchical clustering**

These methods construct the clusters by recursively partitioning the instances in either a top-down or bottom-up manner. These methods can be subdivided as follows [11]

4) **Agglomerative hierarchical clustering**

Every object primarily represents its own cluster. Then clusters are successively merged till the desired cluster structure is obtained.

5) **Divisive hierarchical clustering**

All objects initially belong to single cluster. Then the cluster gets divided into sub-clusters, which successively gets divided into their sub-clusters. This process continues till the desired cluster structure is obtained.

6) **Partitioning clustering**

Partitioning methods relocate instances by moving them from one cluster to another, starting from an initial partitioning. The following subsections present various types of partitioning methods.

7) **K-Means clustering**

The algorithm starts with an initial set of cluster centers, randomly chosen or according to some experimental procedure. In each iteration, each instance is assigned to its nearest cluster center according to the Euclidean distance between the two. Then the cluster centers are re-calculated.

8) **Graph-Theoretic clustering**

Graph theoretic methods are those methods which produce clusters via graphs. The edges of the graph join the instances which are represented as nodes. A graph-theoretic algorithm, which is well known, is based on the Minimal Spanning Tree.

9) **Density-Based methods**

Assume that the points that belong to each and every cluster are drawn from a specific probability distribution (Banfield and Raftery, 1993). The full distribution of the data is assumed to be a mixture of several distributions. The objective of these methods is to identify the clusters and their distribution parameters.

10) **Model-Based clustering methods**

These methods try to optimize the fit between the given data and some mathematical models. Unlike conventional clustering, that recognize groups of objects, model-based clustering methods also find characteristic descriptions for each group, where each and every group represents a concept or class. The most commonly used induction methods are decision trees and neural networks.

11) **Grid-Based methods**

These methods partition the space into a limited number of cells that form a grid structure on which all of the operations for clustering are executed. The main benefit of the approach is its fast processing time.

II. **LITERATURE SURVEY**

According to Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel and Bobby Bhattacharjee [1], the popularity of online social networking sites has given an opportunity to study the properties of online social network graphs at large scale. These graphs help importantly to improve current systems and to develop new applications of OSNs. In this work study of large-scale measurement and examination of the structure of many OSNs done.

According to Stephen J. Hardiman and Liran Katzir [2], the clustering coefficient, a classic measure of network connectivity, are of two types, global and network average. Efficient algorithms for estimating these measures which assume no prior knowledge about...
the network; and access the network using only the publicly available interface. More precisely, this work provides three new estimation algorithms: 1) the first external access algorithm for estimating the global clustering coefficient, 2) an external access algorithm that improves on the accuracy of previous network average clustering coefficient estimation algorithms and 3) an improved external access network size estimation algorithm. The main insight is that only a relatively small number of public interface calls are required to allow algorithms to achieve high accuracy estimation.

According to Liran Katzir, Edo Liberty, Oren Somekh and Ioana A. Cosma [3], algorithms for the number of users estimation of online social networks is present. The proposed algorithms can also estimate the cardinality of network sub-populations. The number of such interactions is strictly limited due to obvious traffic and privacy concerns. Therefore it needs to minimize the number of API interactions for producing good size estimates. Random walk based node sampling is performed in order to adopt the standard abstraction of social networks as undirected graphs.

According to Yong-Yeol Ahn, Seungyeop Han, Haewoon Kwak, Young-Ho Eom, Sue Moon, and Hawoong Jeong [4], OSNs are a fast-emergent business in the Internet. In real-life social networks one can't judge the online relationships and their growth patterns are similar. Three OSN services i.e. MySpace, Cyworld and Orkut reveal the comparative study which consists of more than 10 million users, respectively. Complete data of Cyworld’s friend relationships is accessed and its degree of distribution, clustering property, correlation, and development over time are analyzed.

According to Han Hee Song, Tae Won Cho, Vacha Dave, Yin Zhang, and Lili Qiu [5], proximity deals with the closeness or likeness between nodes in a social network which forms the basis of a range of applications like in social science, information technology, business computer networks, and cyber security challenges to estimate proximity measures in OSNs due to their massive scale and dynamic nature. To overcome this challenge, two unique procedures to powerfully and precisely approximate a large family of proximity measures is developed and also propose a incremental update algorithm to allow near real-time proximity estimation in highly dynamic OSNs. Estimation is done on a huge amount of data collected in five popular online social networks. Link prediction technique is used for proximity estimation.

According to Mark S. Handcock, Adrian E. Raftery and Jeremy M. Tantrum [6], two-stage maximum likelihood method and a Bayesian MCMC method are proposed; the former is faster and simpler, but the latter performs better. Bayesian approximate conditional Bayes factors are also proposed to determine the number of clusters. The model makes it rather easy to execute realistic networks with clustering, potentially helpful as inputs to models of multifaceted systems of which the network is part, such as contagion models of contagious diseases.

III. PROPOSED WORK

The objective of the work is to efficiently estimate the load of online social networking sites. To fulfill this objective, Clustering techniques are used to make the implementation of the approach fast and accurate.

The above flowchart of proposed work Fig. 3 shows that system can be divided into two phases:

A. Preprocessing Phase

Datasets: Data is gathered for three popular social networking sites: Facebook, Twitter and LinkedIn from Internet Libraries. Datasets contain data of 600 cities across the world in terms of Number of users and response time for 10 random servers of each city respectively.

Dimension reduction algorithm: Algorithm is designed to reduce the datasets in order to obtain correct data. Following are the three conditions that are defined in algorithm to attain the reduction of datasets.

- Duplicate rows have to be deleted.
- Empty cell need to be deleted.
- In case, whose server’s average is less than the specified value entered by user is to be deleted.

Algorithm returns the datasets as mentioned in the Table II.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Number of Cities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facebook</td>
<td>400</td>
</tr>
<tr>
<td>Twitter</td>
<td>389</td>
</tr>
<tr>
<td>LinkedIn</td>
<td>400</td>
</tr>
</tbody>
</table>

Algorithm Design

Start
Declare Variables R1, N1, V1, Rag1, i, j, k, ss, m, s, k1, k2, k3, r, s2, p
read value r from user
read value ss from user
Delete duplicate rows

initialize variable p ← 0
initialize variable k1 ← 0
for i ← 1 to r do
    for j ← i+1 to r do
        s ← 0
        for m ← 2 to 11 do
            if value(cells(i,m)) equals value(cells(j,m))
                s ← 1
            end if
        next m
        if value s is 0
            delete.rows(j)
        end if
    next j
next i

Delete rows having empty cells

initialize r ← r-k1
initialize k2 ← 0
initialize i ← 2
Do while i less than or equal to r
s ← 0
for m ← 2 to 11
    If cells (i,m) equals 0 then
        s ← 1
    end if
next m
If s equals 1 then
    row(i).delete
    k2 ← k2+1
else
    i ← i+1
end if
end loop

Delete rows whose average is less than the specified value entered by user

initialize p ← 1
initialize k3 ← 0
initialize i ← 2
Do while i is less than or equal to r
s ← 0
if value of (cells(i,12)) less than ss then
    row(i).delete
    r ← r-1
    k3 ← k3+1
else
    i ← i+1
end if
end loop

B. Query Phase

K-Means and Agglomerative Hierarchical clustering techniques are applied on reduced datasets of Facebook, Twitter and LinkedIn using XLStat.

K-Means clustering: This Clustering Technique is applied on datasets at K=5, 10, 15, and 20 (Where K = Number of Clusters) to form the clusters of different cities in order to identify the load of the social networking sites (Facebook, Twitter and LinkedIn).

K-Means clustering algorithm: It assumes that we know the number of clusters k. This is an iterative algorithm which keeps track of the cluster centers (means). The centers are in the same feature space as x.

1. Choose k centers μ_1, ..., μ_k randomly.
2. Repeat
3. Assign x_1,...x_n to their closest centers, respectively.
4. Update μ_i to the mean of the items assigned to it.
5. until the clusters no longer change.

Agglomerative hierarchical clustering: This Clustering Technique is also applied on datasets at K=5, 10, 15, and 20 (Where K = Number of Clusters) to form the clusters of different cities in order to identify the load of the social networking sites (Facebook, Twitter and LinkedIn).

AHC Algorithm: This is a very simple procedure:
1. Initially each item x_1, ..., x_n is in own cluster C_1, C_2, C_3, ..., C_n.
2. Repeat until there is only single cluster left.
3. Merge the nearest clusters, say C_i and C_j.
4. The result is a cluster tree. One can cut the tree at any level to produce different clustering. A little thought reveals that “the nearest clusters” are not well-defined, since we only have a distance measure d(x, x_0) between items. This is where the variations come in:
   • d(C_i,C_j) = min_xєC_i,x_0єC_j d(x, x_0). This is known as single-linkage. It is equivalent to the minimum spanning tree algorithm. Anyone can set a threshold and stop clustering once the distance between clusters is above the threshold. Single-linkage tends to produce large and skinny clusters.
   • d(C_i,C_j) = max_xєC_i,x_0єC_j d(x, x_0). This is known as complete-linkage. Clusters tend to be compact and roughly equal in diameter.
   • d(C_i,C_j) = \frac{\sum_{xєC_i \cap C_j} d(x,x_0)}{|C_i|·|C_j|}. This is the average distance between items somewhere between single-linkage and complete-linkage.

After applying above mentioned techniques on reduced datasets of online social networking sites using XLStat statistical tool, we obtained various clusters to estimate load of various OSNs.

IV. EXPERIMENTAL SETUP

XLSTAT is an extendible toolkit for data analysis and statistical software, discovery and exploration implemented as an add-in to the Microsoft Excel 2007/11 software [12] is used. Two Clustering techniques are used to estimate the load of social networking sites using XLStat.

• Agglomerative Hierarchical Clustering
• K-Means Clustering
V. EXPERIMENT RESULT

A. For K=5 (Where K = Number of Clusters) of K-Means Clustering

Fig. 4 shows the results of dataset at K=5 according to K-Means as per Table III for Facebook, Twitter & LinkedIn i.e. Maximum number of cities are contained in Cluster-2 while Minimum number of cities are contained in Cluster-5.

B. For K=10 (Where K = Number of Clusters) of K-Means Clustering

Fig. 5 shows the results of dataset at K=10 according to K-Means as per Table IV for Facebook, Twitter & LinkedIn i.e. Maximum number of cities are contained in Cluster-2 while Minimum number of cities are contained in Cluster-10.

C. For K=5 (Where K = Number of Clusters) of Agglomerative Hierarchical Clustering

Fig. 6 shows the results of dataset at K=5 according to Agglomerative Hierarchical Clustering Technique as per Table V for Facebook, Twitter & LinkedIn i.e. Maximum number of cities are contained in Cluster-2 while Minimum number of cities are contained in Cluster-5.

D. For K=10 (Where K = Number of Clusters) of Agglomerative Hierarchical Clustering

Fig. 7 shows the results of dataset at K=10 according to Agglomerative Hierarchical Clustering Technique as per Table VI for Facebook, Twitter & LinkedIn i.e. Maximum number of cities are contained in Cluster-2 while Minimum number of cities are contained in Cluster-9 & 10.

E. For K=5 (Where K = Number of Clusters) of K-Means Clustering

Fig. 8 shows the results of dataset (response time) at K=5 according to K-Means as per Table VII for Facebook maximum number of cities are contained in Cluster-3 while minimum number of cities are contained in Cluster-1 and for Twitter maximum number of cities are contained in Cluster-2 while minimum number of cities are contained in Cluster-4 & for LinkedIn.

TABLE III. COMPARATIVE RESULT OF DATASETS AT K=5 (K-MEANS)

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
<th>Class 4</th>
<th>Class 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facebook</td>
<td>67</td>
<td>238</td>
<td>54</td>
<td>40</td>
<td>1</td>
</tr>
<tr>
<td>Twitter</td>
<td>51</td>
<td>268</td>
<td>35</td>
<td>33</td>
<td>2</td>
</tr>
<tr>
<td>LinkedIn</td>
<td>54</td>
<td>293</td>
<td>13</td>
<td>39</td>
<td>1</td>
</tr>
</tbody>
</table>

TABLE IV. COMPARATIVE RESULT OF DATASETS AT K=10 (K-MEANS)

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
<th>Class 4</th>
<th>Class 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facebook</td>
<td>48</td>
<td>186</td>
<td>38</td>
<td>49</td>
<td>20</td>
</tr>
<tr>
<td>Twitter</td>
<td>28</td>
<td>194</td>
<td>24</td>
<td>57</td>
<td>26</td>
</tr>
<tr>
<td>LinkedIn</td>
<td>19</td>
<td>218</td>
<td>15</td>
<td>59</td>
<td>11</td>
</tr>
<tr>
<td>Datasets</td>
<td>Class 6</td>
<td>Class 7</td>
<td>Class 8</td>
<td>Class 9</td>
<td>Class 10</td>
</tr>
<tr>
<td>Facebook</td>
<td>31</td>
<td>4</td>
<td>6</td>
<td>17</td>
<td>1</td>
</tr>
<tr>
<td>Twitter</td>
<td>13</td>
<td>25</td>
<td>19</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>LinkedIn</td>
<td>21</td>
<td>4</td>
<td>27</td>
<td>25</td>
<td>1</td>
</tr>
</tbody>
</table>

TABLE V. COMPARATIVE RESULT OF DATASETS AT K=5 (AHC)

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
<th>Class 4</th>
<th>Class 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facebook</td>
<td>55</td>
<td>292</td>
<td>49</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Twitter</td>
<td>30</td>
<td>278</td>
<td>42</td>
<td>37</td>
<td>2</td>
</tr>
<tr>
<td>LinkedIn</td>
<td>65</td>
<td>293</td>
<td>12</td>
<td>4</td>
<td>26</td>
</tr>
</tbody>
</table>

TABLE VI. COMPARATIVE RESULT OF DATASETS AT K=10 (AHC)

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
<th>Class 4</th>
<th>Class 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facebook</td>
<td>26</td>
<td>221</td>
<td>33</td>
<td>71</td>
<td>26</td>
</tr>
<tr>
<td>Twitter</td>
<td>29</td>
<td>278</td>
<td>29</td>
<td>16</td>
<td>13</td>
</tr>
<tr>
<td>LinkedIn</td>
<td>22</td>
<td>293</td>
<td>15</td>
<td>12</td>
<td>3</td>
</tr>
</tbody>
</table>

TABLE VII. COMPARATIVE RESULT OF DATASETS AT K=5 (K-MEANS)

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
<th>Class 4</th>
<th>Class 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facebook</td>
<td>16</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Twitter</td>
<td>18</td>
<td>27</td>
<td>8</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
maximum number of cities are contained in Cluster-1 while minimum number of cities are contained in Cluster-3.

**TABLE VII. COMPARATIVE RESULT OF DATASETS (RESPONSE TIME) AT K=5 (K-MEANS)**

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
<th>Class 4</th>
<th>Class 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facebook_RT</td>
<td>68</td>
<td>80</td>
<td>91</td>
<td>82</td>
<td>79</td>
</tr>
<tr>
<td>Twitter_RT</td>
<td>83</td>
<td>86</td>
<td>76</td>
<td>61</td>
<td>83</td>
</tr>
<tr>
<td>LinkedIn_RT</td>
<td>93</td>
<td>74</td>
<td>68</td>
<td>86</td>
<td>79</td>
</tr>
</tbody>
</table>

**F. For K=10 (Where K = Number of Clusters) of K-Means Clustering**

Fig. 9 shows the results of dataset (response time) at K=10 according to K-Means as per Table VIII for Facebook maximum number of cities are contained in Cluster-1 while minimum number of cities are contained in Cluster-2 and for Twitter maximum number of cities are contained in Cluster-6 while minimum number of cities are contained in Cluster-7 & for LinkedIn maximum number of cities are contained in Cluster-1 while minimum number of cities are contained in Cluster-10.

**TABLE VIII. COMPARATIVE RESULT OF DATASETS (RESPONSE TIME) AT K=10 (K-MEANS)**

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
<th>Class 4</th>
<th>Class 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facebook_RT</td>
<td>49</td>
<td>44</td>
<td>45</td>
<td>37</td>
<td>32</td>
</tr>
<tr>
<td>Twitter_RT</td>
<td>38</td>
<td>36</td>
<td>42</td>
<td>44</td>
<td>33</td>
</tr>
<tr>
<td>LinkedIn_RT</td>
<td>52</td>
<td>34</td>
<td>41</td>
<td>44</td>
<td>43</td>
</tr>
</tbody>
</table>

**G. For K=5 (Where K = Number of Clusters) of Agglomerative Hierarchical Clustering**

Fig. 10 shows the results of dataset (response time) at K=5 according to Agglomerative Hierarchical Clustering Technique as per Table IX for Facebook maximum number of cities are contained in Cluster-3 while minimum number of cities are contained in Cluster-4 and for Twitter maximum number of cities are contained in Cluster-3 while minimum number of cities are contained in Cluster-4 & for LinkedIn maximum number of cities are contained in Cluster-2 while minimum number of cities are contained in Cluster-5.

**TABLE IX. COMPARATIVE RESULT OF DATASETS (RESPONSE TIME) AT K=5 (AHC)**

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
<th>Class 4</th>
<th>Class 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facebook_RT</td>
<td>81</td>
<td>65</td>
<td>102</td>
<td>52</td>
<td>100</td>
</tr>
<tr>
<td>Twitter_RT</td>
<td>85</td>
<td>116</td>
<td>62</td>
<td>64</td>
<td>62</td>
</tr>
<tr>
<td>LinkedIn_RT</td>
<td>99</td>
<td>113</td>
<td>61</td>
<td>69</td>
<td>58</td>
</tr>
</tbody>
</table>

**H. For K=10 (Where K = Number of Clusters) of Agglomerative Hierarchical Clustering**

Fig. 11 shows the results of dataset (response time) at K=10 according to Agglomerative Hierarchical Clustering Technique as per Table X for Facebook maximum number of cities are contained in Cluster-3 while minimum number of cities are contained in Cluster-6 and for Twitter maximum number of cities are contained in Cluster-8 while minimum number of cities are contained in Cluster-5 & for LinkedIn maximum number of cities are contained in Cluster-2 while minimum number of cities are contained in Cluster-10.
and LinkedIn) at various values of load of three social networking sites (Facebook, Twitter, LinkedIn) by estimating the clustering techniques.

The observations concluded two parameters i.e. Number of Users and Response Time by estimating the load of three social networking sites (Facebook, Twitter, LinkedIn) at various values of ‘K’ (Where K = Number of clusters) of both the techniques K-means and Agglomerative Clustering. The results show the clusters which arise from both the techniques contain various numbers of objects. Therefore, this directs us to jump into the conclusion that all the objects which come under that particular cluster cover same load.

The next challenge is to improve load estimation techniques by including more number of server specification to show the effectiveness of results. This work also be extended to balance the sever load and improve the response time of servers by applying few more clustering algorithms.

TABLE X. COMPARATIVE RESULT OF DATASETS (RESPONSE TIME) AT K=10 (AHC)

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
<th>Class 4</th>
<th>Class 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facebook_RT</td>
<td>49</td>
<td>65</td>
<td>66</td>
<td>19</td>
<td>52</td>
</tr>
<tr>
<td>Twitter_RT</td>
<td>53</td>
<td>48</td>
<td>40</td>
<td>32</td>
<td>21</td>
</tr>
<tr>
<td>LinkedIn_RT</td>
<td>34</td>
<td>72</td>
<td>61</td>
<td>43</td>
<td>26</td>
</tr>
</tbody>
</table>

VI. CONCLUSION

The expansion of OSNs has increased significantly in the recent years, counting more than 1.2 billion users in 2014. Unlike previous web applications, OSNs are user centered and offer a various tools to smooth the progress of information sharing and communication between their users. Study is inspired by the fact that the load of OSNs are increasing rapidly in terms of many number of users which pushed us further to analyze the online population making use of OSNs. Therefore, we have estimated the load of OSNs by forming the clusters of datasets using two clustering techniques K-Means and Agglomerative clustering techniques.

ACKNOWLEDGMENT

Gratitude is expressed to the editor and the anonymous reviewers for their valuable and constructive comments, which will be very helpful in improving the quality of the paper.

REFERENCES


Deepthi Bhagwani was born in Gwalior, M.P., India. She graduated in Science (3-year university degree, PGV College of Science, Gwalior, M.P., India, 2004), Master in Computer Application (Institute of Technology & Management, Gwalior, M.P., India, 2007), and she completed her Master of Technology in Computer Technology & Application ( Technocrat Institute of Technology, Bhopal, M.P., India, 2014). This was the dissertation work in Load Estimation of Social Networking Services through Clustering Techniques during her Master of Technology in Computer Technology & Application. Her research interests include clustering techniques, data mining.