Detecting the Age of a Person Through Web Browsing Patterns: A Review

Chinmay Swami, Prasad Tarte, Sagar Rakshe, Sumit Raut, Nuzhat F. Shaikh

Abstract*—*As the use of internet is growing day by day, the basic attributes of the user such as his age, his location, his preferences are of a great value to various business corporations. We are aware of the fact that there is some connection between the browsing behavior and the basic characteristics of the user. In this paper we have made an effort to summarize the various aspects related to detecting the age of the person through his browsing patterns.

Index Terms*—*Age Prediction, Back-Propagation Neural Networks, Connectionism and neural nets, Concept learning.

I. INTRODUCTION

Various organizations that have their business depended on the internet have began paying more attention towards providing modified service in order to increase the user experience. Google, various online shopping sites, Yahoo etc are some perfect examples of it. Various online shopping sites keep track of their user's preferences and use this information to provide the user a better browsing experience. For example, shopping sites such as Flipkart, Snapdeal etc keep track of what all the user is searching for and after recognizing the requirement of user they provide various suggestions for the user as well as the provide details of his previously searched items. Google Personal arranges user's search outcome according to their search records including their earlier search results and news headings checked by them [5].Due to the increase in E –Commerce there has also been an significant increase in online advertising. Behaviour targeting is a popular technique used in them. Behaviour targeting consists of various techniques that are used by the online advertisers aimed at increasing the effectiveness of the advertisements with the help of user web-browsing behaviour data. Hence, the users web-browsing behaviour plays an important role in the today's internet age. Also there are various aspects of a website that is not appropriate for kids, but controlling what kids see or do over the web is a tedious task. Also, due to significant growth in cyber crime over the last couple of year's people tend to be more secretive about their web activities and refrain from disclosing any of their private information. There are various methods available to predict user's demographic attributes one of the approach to find the age of the person using his browsing pattern is by using one of the techniques from artificial neural network called as multilayer perception(MLP) with back propagation algorithm [1].There are various machine learning approaches used to predict the demographic information of website using the information generated by considering various aspects of the website and not relying on any information that is obtained from any kind of survey performed by user's [2].
As per survey conducted by "Statista" in June 2014

Age Group	Internet Usage %
15-24	26.7
25-34	26.6
35-44	20.3
45-54	13.7
55+	12.7

Table 1: Internet usage statistics

Hence having knowledge about the age of the user can be quite handy as the online advertise companies can provide advertisements according to the age group of user. For example, if the user is 55+, they mostly are interested in religious things such as travelling to religious parts of the country, reading religious books etc. So the advertisements agencies will target the user with more religious kind of advertisements. Also various Natural language processing techniques can be used to predict the age of the person [3].
In this paper we have discussed various methodologies to successfully detect the age of the user based on his browsing patterns and the accuracy of the various such methodologies. The remaining paper is divided into 6 different sections. In section 2, Age prediction problem definition is stated. In section 3, various solutions that are present are reviewed. Section 4 and 5 contains the conclusion and future scope respectively.

II. PROBLEM DEFINITION

Here we are trying to predict the age of the user based on his web browsing patterns. Here the age predictions are done in terms of age group.

Age Group	Age
Teenager	<=24
Young Adult	25 to 35
Adult	35+

Table 2: Age Group (can vary)

Basically the browsing history is a set that consists of various elements like the URL, Time Request, Time Response, Time of Completion etc. The system will have a previously defined data set that contains all the data mentioned above. Also the system will keep on collecting

Manuscript Received on November, 2014.

Chinmay Swami, Computer Engineering, University of Pune, Modern education society college of engineering, Pune, India .
Prasad Tarte, Computer Engineering, University of Pune, Modern education society college of engineering, Pune, India .
Sagar Rakshe, Computer Engineering, University of Pune, Modern education society college of engineering, Pune, India .
Sumit Raut, Computer Engineering, University of Pune, Modern education society college of engineering, Pune, India .
Nuzhat F. Shaikh, Associate Professor in the Department of Computer Engineering at MES College of Engineering, Pune, India.

all the above mentioned data while the user is browsing the web and keep on storing it into a local database. After collecting sufficient amount of data the system will match the recently collected data will the previously present data and will predict the age group of the user.

III. VARIOUS AGE PREDICTION METHODS

A. Artificial Neural Network

Here the age prediction problem is tackled with the help of artificial neural network. An ANN is a mathematical representation of the human neural architecture, representing its “learning” and “generalization” capabilities. Neural networks basically are a system of interconnected neurons which consists of 3 parts (as shown in figure 1) they are:

A. Input Layer
B. Hidden Layers
C. Output Layer

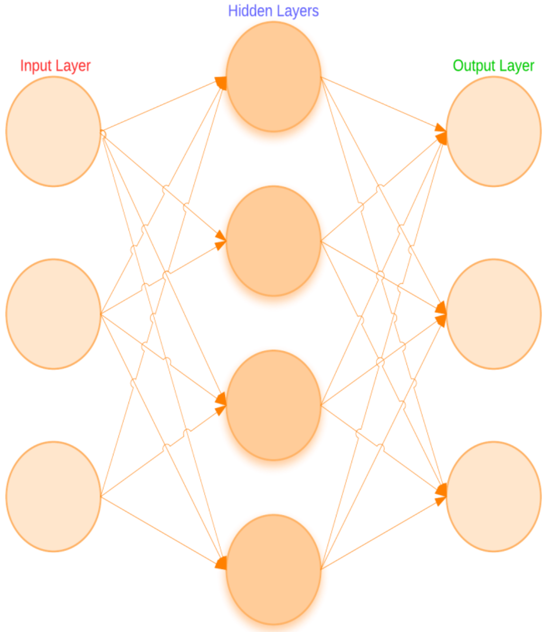


Figure 1: Basic Structure of ANN

Feed forward neural network is one of the first and simplest of neural network where information only moves in one direction i.e. forward and contains no cycles and loops. There are two types of Feed forward neural networks:

A. Single-Layer Perception
B. Multi-Layer Perception

Multi-layer perception is fully connected i.e. output from each neuron either input or hidden is distributed to every other neuron in next layer. The number of hidden layers may vary. Back-Propagation is one of the most popular learning method used in multi-layer networks. In this method the output value is compared with some predefined correct value to calculate an appropriate value of some predefined error function. This computed value of error function is then fed back through the network. With the help of this information the algorithm alters the weights of each connection between neuron with the intension of reducing the value of error function. This process is repeated continuously until the value of error function is quite small. At that instance we can say that the network has learned certain function. Number of Hidden layers in the network plays an important role in accuracy for the network. There is no fix law for calculating the number of hidden layers a neural network should have. There are some empirically-derived rules-of-thumb, of these, the most commonly relied on is 'the optimal size of the hidden layer is usually between the size of the input and size of the output layers' stated by Jeff Heaton [4].

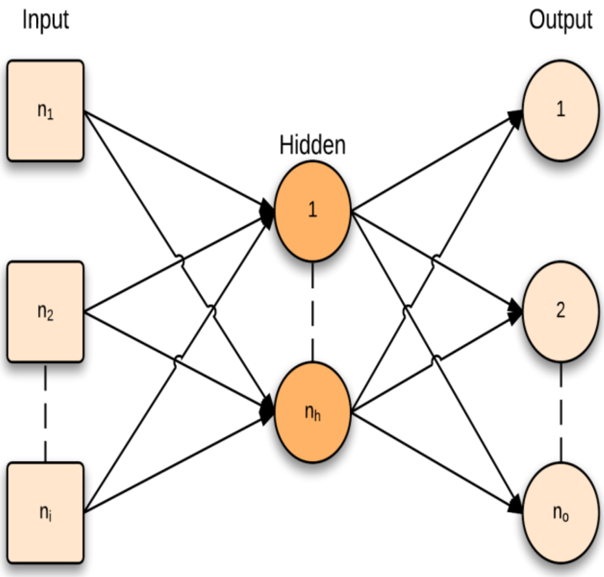


Figure 2: Schema of a Back-Propagation Neural Network (BPNN)

In Figure 2 the BPNN has 3 layers in which all the node in input layer are connected to ever other node in the hidden layer and all the nodes in hidden layer are connected to every other node in next layer (Hidden or output). All the connections between nodes are directed and no nodes from the same layer are connected. Each of these connections have weights and these weights are altered using the BPNN algorithm in the training (learning) process. Also it is important to choose the number of training samples. According to Baum and Haussler [6], to guarantee a level of performance on a set of test samples obtained from the same training data we can place a bound on the training samples.

B. Apriori Algorithm

Apriori Algorithm is one of the most influential algorithms for mining frequent item set for Boolean association rules and association rule learning over the database. In short Apriori is a classic algorithm used in data mining for learning association rules. Apriori is designed to operate on databases which consist of transactions. Each transition is viewed as a set of items (itemset).

Key concepts in Apriori algorithms are:-

A. Frequent Itemsets:
The set of items with minimum support(Denoted as L_i for i^{th}itemset)
B. Apriori Property:
Any sub set of frequent itemset must be a frequent itemset i.e. if{AB} is a frequent itemset then both {A},{B} must be frequent itemset.
C. Join Operation:
To find L_k, a set of candidate K-itemsets is generated by joining L_{k-1} with itself.
Prune Step: Any non frequent (k-1)-itemset cannot be subset of frequent k-itemset
Pseudo Code:

```
$C_k$: Candidate itemset of size k
$L_k$: frequent itemset of size k
$L_1$= {frequent items};
for(k= 1; $L_k$!=ϕ; k++) do begin
$C_{k+1}$= candidates generated from $L_k$;
for each transaction tin database do
increment the count of all candidates in $C_{k+1}$ that are contained in t
    $L_{k+1}$= candidates in $C_{k+1}$with min_support
    end
    return $U_k L_k$;
```

There are various methods available for improving the efficiency of Apriori algorithm. They are:

A. Hash based Itemset Counting
B. Transaction Reduction
C. Partitioning
D. Sampling
E. Dynamic Itemset Counting

C. K-Means Algorithm

K-Means also known as K-Means Clustering is one of the popular clustering analysis algorithm in data mining originally derived from Digital signal Processing and is generally categorized as a method of vector quantization. It's also been applied in machine learning field. The main objective of K-Means is to partition n observations in k clusters where each observation belongs to cluster with nearest mean. The results obtained are of data space into Voronoi cells.

In general, Voronoi diagrams are used to divide space into number of regions as shown in figure 3. These regions are termed as Voronoi cells.

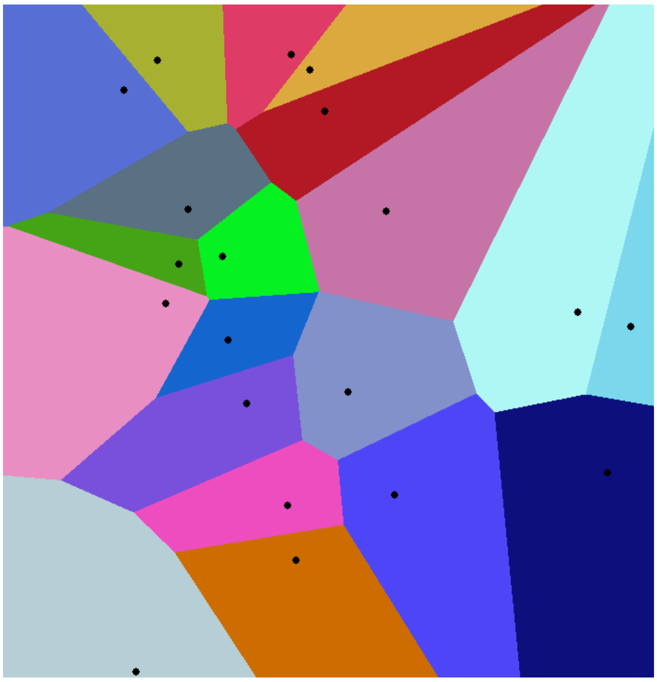


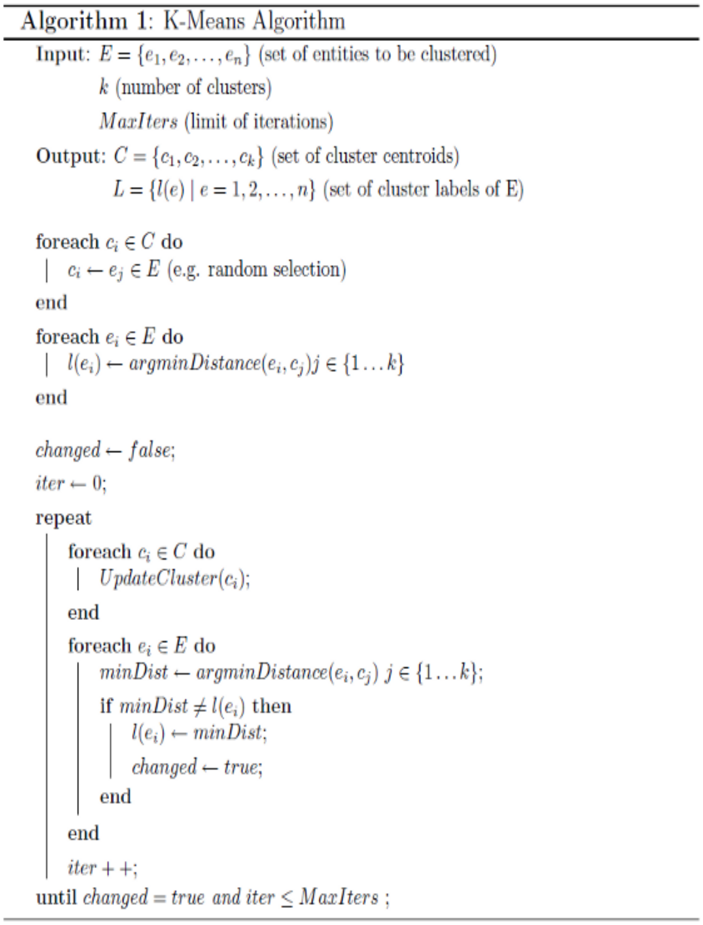
Figure 3: Euclidean distance Voronoi Diagram.

K-means is a simple procedure and follows a convenient way to classify the given dataset through some clusters (k) a priori. The main idea over here is to define k centroids, one for each cluster. It's important to place these centroids at appropriate location as the end result depends on it. Hence, it's preferred that the centroids are placed as far as possible from each other. In the next step each point is associated to the nearest centroids where these points belong to a given dataset. If there are no points remaining then an early grouping is done and we computer the k new centroids and these newly computed centroids do not move i.e. no more changes are performed.

Finally the algorithm aims at minimizing the objective function which is,

$$J = \sum_{j-1}^{K} \sum_{i-1}^{n} ||x_i^{(j)} - c_j||^2 \text{ ---(1)}$$

Where, $||x_i^{(i)} - c_j||^2$ is the chosen distance between data point $x_i^{(i)}$ and cluster centre c_j, and indicates the distance of n data points from their respective cluster centres.



```
Algorithm 1: K-Means Algorithm
Input: $E = \{e_1, e_2, \dots, e_n\}$ (set of entities to be clustered)
       $k$ (number of clusters)
       $MaxIters$ (limit of iterations)
Output: $C = \{c_1, c_2, \dots, c_k\}$ (set of cluster centroids)
        $L = \{l(e) \mid e = 1, 2, \dots, n\}$ (set of cluster labels of E)

foreach $c_i \in C$ do
    $c_i \leftarrow e_j \in E$ (e.g. random selection)
end
foreach $e_i \in E$ do
    $l(e_i) \leftarrow argminDistance(e_i, c_j) j \in \{1 \dots k\}$
end

$changed \leftarrow false$;
$iter \leftarrow 0$;
repeat
    foreach $c_i \in C$ do
        $UpdateCluster(c_i)$;
    end
    foreach $e_i \in E$ do
        $minDist \leftarrow argminDistance(e_i, c_j)\ j \in \{1 \dots k\}$;
        if $minDist \neq l(e_i)$ then
            $l(e_i) \leftarrow minDist$;
            $changed \leftarrow true$;
        end
    end
    $iter++$;
until $changed = true$ and $iter \leq MaxIters$ ;
```

[Note that the last line of the pseudo code should be "until changed = false or iter>MaxIters;".]
The Algorithm comprises of following steps

1] Place K points into the space represented by the objects that are being clustered. The initial group centroids are represented by these points.
2] Each objects is assigned to the group having closest centroid .
3] After assigning all objects, recalculate the positions of the K centroids.
4] Repeat Steps 2 and 3 until the centroids can no longer move. Hence, a separation of the objects into groups is produced from which the metric to be minimized can be calculated.

Some of the drawbacks faced by K- means are firstly, the results completely depend on value k. Also it’s quite difficult to compare the quality of clusters produced. If there are fixed number of clusters then it’s quite difficult to predict the value of k. K-means doesn’t work with non-Globular clusters.

Finally we summarize K-Means by saying that it can be viewed as a greedy algorithm to partition n samples into k clusters to minimize the sum of squared distances to cluster centres.

D. ID3

Data mining is used to extract the required data from large databases. The data mining algorithm is the mechanism that creates mining models. To create a model, an algorithm first learns the rules from a set of data then looks for specific required patterns and trends according to those rules. ID3 is a simple decision tree algorithm developed by Ross Quinlan (1983). The basic idea of ID3 algorithm is to create a decision tree of given set, by using top-down greedy search to check each attribute at every tree node. To select the most useful attribute using classification technique, we present a metric information gain and to catch an optimal way to classify an educated set, we need to minimize the depth of the tree

The ID3 algorithm is used to build a decision tree, given a set of non-categorical attributes C_1,C_2, .., Cn, the categorical attribute C, and a training set T of records.

Algorithm of ID3 is as follows :

```
1) Function ID3 (R: a set of non-categorical attributes,
2) C: the categorical attribute,
3) S: a training set) returns a decision tree;
4) begin
5) If S is empty, return a single node with value Failure;
6) If S consists of records all with the same value for the
categorical attribute,
7) Return a single node with that value;
8) If R is empty, then return a single node with as value
the most frequent of the values of the categorical attribute
that are found in records of S;
9) Let D be the attribute with largest Gain (D,S)
10) Among attributes in R;
11) Let {dj| j=1,2, .., m} be the values of attribute D;
12) Let {Sj| j=1,2, .., m} be the subsets of S consisting
respectively of records with value dj for   attribute D;
13) Return a tree with root labeled D and arcs labeled.
14) d1, d2, .., dm going respectively to the trees
15) ID3(R-{D}, C, S1), ID3(R-{D}, C, S2), ..,
ID3(R-{D}, C, Sm);
16) end ID3;
```

E. Clustering

Itis the process of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar to each other than to those in other groups. It is a main task for data mining, and a common technique for statistical data analysis ,used in many fields, including machine learning, pattern recognition, image analysis, data retrieval, and bioinformatics. Cluster analysis is not one specific algorithm, but it is the task to be solved. It is the combination of various algorithms that differs from each other. The appropriate clustering algorithm and parameter settings (including values such as the distance function to use) depend on the individual data set and intended use of the results. Cluster analysis is not an automatic task. It is necessary to modify the data pre-processing and model parameters until the desired result is achieved.

There are many types of algorithms present for the clustering as the clustering algorithm itself is a combination of various different algorithms. The various algorithms are as follows:

A] Connectivity based clustering (hierarchical clustering)

It is also known as hierarchical clustering, is based on the ideology of objects being more related to closer objects than to the objects that are farther away. This algorithm mainly connects the "objects" to "clusters" depending on the distance between them. The clusters are described based on the maximum distance needed to connect parts of the cluster.

As shown in figure 4, At 35 clusters, the biggest cluster are fragmented into smaller parts, while before it was still connected to the second largest due to the single-link Effect. At 20 clusters clustering is re-applied and we get the output as shown in second part of figure 4

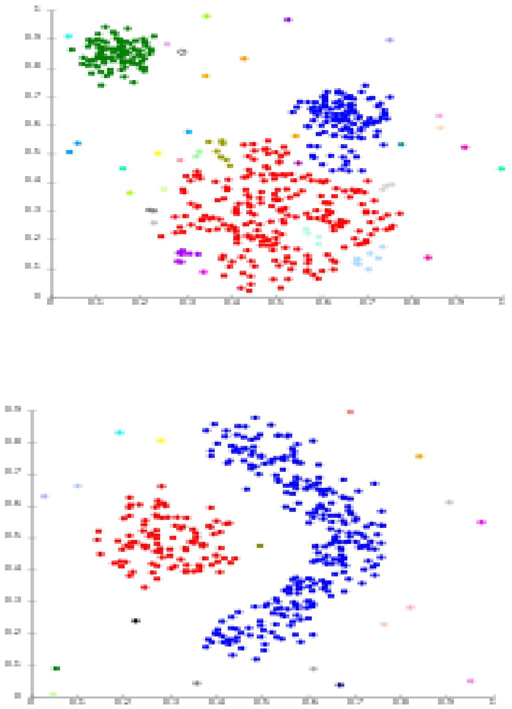


Figure 4: Output in the form of Voronoi diagram

B] Centroid-based clustering

In centroid-based clustering, represents clusters by a central vector, which may or may not compulsorily be a member of the data set. K-means clustering gives a formal definition as an optimization problem when the number of clusters is fixed to k.

C] Distribution-based clustering

Distribution-based clustering is a semantically strong method, as it not only provide clusters, but also produces complex models for the clusters that can also capture correlation and dependence of attributes. However, a drawback of using these algorithms is that it puts an extra burden on the user: of choosing an appropriate data model to optimize, and there exist many real data sets, which do not have mathematical model available the algorithm is able to optimize.

D] Density-based clustering

In density-based clustering, clusters are defined as areas of higher density than the remainder of the data set. Objects in these sparse areas - are usually considered to be noise and border points if required to separate clusters.

One of the most widely used density based clustering method is DBSCAN. As opposed to various newer methods, it features a well-defined cluster model called "density-reachability". A cluster is made up of all density-connected objects (forming a cluster of an arbitrary shape) as well as all objects that are within these objects' range. Another interesting property of DBSCAN is has a comparatively low level of complexity .It requires a linear number of range queries on the database and that it will discover essentially the same results.

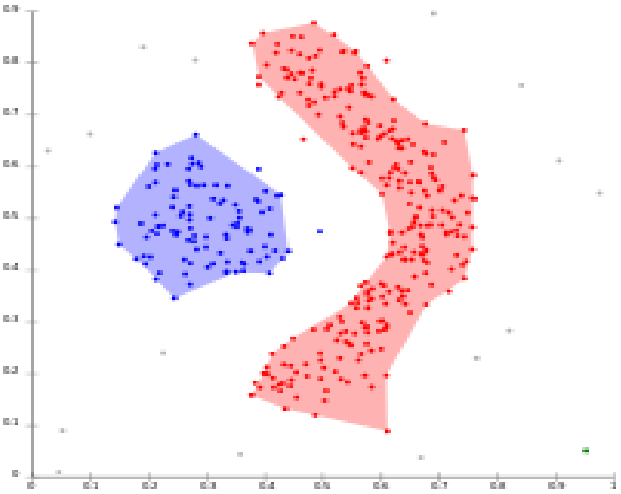


Figure 5: Density-based clustering with DBSCAN

Various advancements have occurred in the field of clustering algorithm. In past few years considerable amount of effort has been put into increasing the effectiveness of algorithm of the existing algorithms. With the recent need to process larger data sets (widely known as big data) the approach towards semantic meaning of the generated clusters for performance has been increasing. This caused the development of pre-clustering methods such as canopy clustering, which processes the huge data sets effectively, but the resulting "clusters" are merely a rough pre-partitioning of the data set to then analyze the partitions with existing slower methods such as k-means clustering.

IV. CONCLUSION

This paper focuses on various methodologies that can be used to predict the age of the user based on his internet browsing pattern. Just like the coin has two sides the methodologies stated above also have its pro's and con's. Multi-layer perception Feedforward algorithm of Neural Network seems to tackle the problem definition more efficiently, but if more of statistical data analysis is required then Apriori algorithm fits the best.

V. FUTURE SCOPE

Nowadays various people have fake profiles on various social sites, our proposed algorithm can be used to identify such profiles with fake information. Also, as lots and lots of kids are using the internet, the system can be used to implement parental control. In the current work we predict only age of user, but we can extend it to identifying various demographic information such as location, gender, occupation etc.

REFERENCES

[1] Misha Kakkar, DivyaUpadhyay, "Web Browsing Behaviors based on age detection", ISSN: 2231-2307, Volume-3, Issue-1, March 2013, International Journal of Soft Computing and Engineering (IJSCE)

[2] *Santosh Kabbu, Eui-Hong Han, George Karypis*, "Content-Based Methods for Predicting Web-Site Demographic Attributes". NSF (IIS-0905220, OCI-1048018,IOS-0820730), NIH (RLM008713A), and the Digital Technology Centreat the University of Minnesota

[3] *Claudia Peersman, Walter Daelemans*, Leona Van Vaerenbergh,, "Predicting Age and Gender in Online Social Networks", Conference'10, Month 1–2, 2010, City, State, Country, Copyright 2010 ACM 1-58113-000-0/00/0010.

[4] Jeff Heaton, Heaton Research, Inc. (25 November 2005),Introduction to neural networks with java,ISBN-10: 097732060X, ISBN-13:978-0977320608.

[5] A. Lenhart, S. Fox. Bloggers: A portrait of the internet's new storytellers. http://www.pewinternet.org/pdfs/PIP%20Bloggers%20Report%20July%2019%202006.pdf

[6] E.B. Baum and D. Haussler, ``What size net gives valid generalization?,'' *Neural Computation*, vol. 1, no. 1, pp. 151-160, 1989.

[7] M.H Hassoun "Fundamentals of Artificial Neural Networks", IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 42, NO. 4, JULY 1996.

[8] Jonathan Schler, Moshe Koppel, Shlomo Argamon,James Pennebaker "Effects of Age and Gender on Blogging" Copyright © 2005, American Association for Artificial Intelligence(www.aaai.org)

[9] Reyhaneh Tamimi, Prof. Dr. Mohammad Ebrahim Mohammad pourzarandi, " The Application of Web Usage Mining In E-commerce Security", 978-1-4799-0393-1/13/$31.00 ©2013 IEEE

Chinmay Swami, perusing Computer Engineering, from University of Pune ,Mo dern education society college of engineering, Pune, India

Sagar Rakshe, perusing Computer Engineering, from University of Pune ,Modern education society college of engineering, Pune, India

Sumit Raut, perusing Computer Engineering, from University of Pune ,Modern education society college of engineering, Pune, India

Prasad Tarte perusing Computer Engineering, from University of Pune ,Modern education s ociety college of engineering, Pune, India

Nuzhat F. Shaikh, (Our Guide) She is currently a research scholar working for a Ph. D degree at Nanded university. At present, she is working as an Associate Professor in the Department of Computer Engineering at MES College of Engineering, Pune. She is the reviewer of many international journals and has published various technical papers at conferences and in reputed journals.