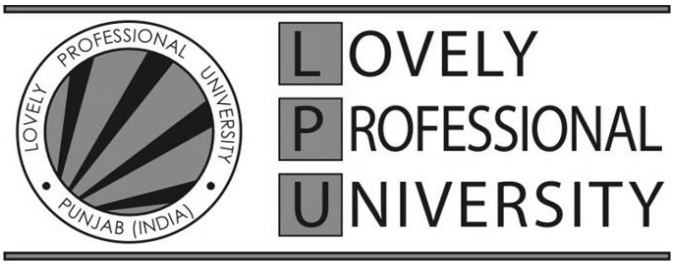

IDENTIFYING EFFECTIVE ATTRIBUTES FOR STRUCTURING SOCIAL GROUPS IN SOCIAL MEDIA USING PRINCIPAL COMPONENT ANALYSIS

Dissertation Proposal submitted

by

Samiya Shafi

to

Department of Computer Science and Engineering

In partial fulfillment of the Requirement for the Award of the Degree of

Master of Technology in Computer Science and Engineering

Under the guidance of

Harshpreet Singh

Asst. Professor

(May 2015)

ABSTRACT

Big Data refers to large amount of data generated from heterogeneous resources. The data generated from social media aims in providing information about the individual and their respective behaviour. The user attributes in the social data also provides a structure to form various social networks. This motivation of the work carried out in this study focuses on analyzing the social data for finding attributes which are effective in making circles and lists in social media. Using various dimensions such as name, school, place, location, birthday, degree, class, school, name, hometown etc the study aims in providing the principle components which aids in building the social network.

The study employs Principal Component Analysis for the analysis of social media data for Circle/List formation. Principal Component Analysis is mathematical procedure to reduce number of dimensions of dataset by maintaining its original variability. The approach is utilized for isolating the attribute effective in making of circles. Interpretation and validation of the proposed methodology is plotted using scree test.

CERTIFICATE

This is to certify that *Samiya Shafi* bearing Registration no. 11307674 has completed M.Tech Dissertation titled, ***"Identifying effective attributes for structuring social groups in Social media using Principal Component Analysis"*** under my guidance and supervision. To the best of my knowledge, the present work is the result of her original investigation, effort and study. No part of the dissertation has ever been submitted for any other degree or diploma at any university.

The dissertation is fit for the submission and the partial fulfillment of the conditions for the award of M.Tech Computer Science & Engineering.

Date:

Signature
Mr. HARSHPREET SINGH
Asst. Professor
School of Computer Science and Engineering
Uid: 17478
Lovely Professional University
Phagwara, Punjab.

ACKNOWLEDGEMENT

I convey my most heartfelt and deepest gratitude to our respected teacher **Mr. HARSHPREET SINGH** for their support, guidance, advice, understanding and supervision throughout this dissertation study. I would also like to thank my family and all the people who provided me with the facilities being required and conductive conditions for my M.Tech dissertation work.

I would also express my warm thanks for everyone who supported me throughout the course of **M.Tech Dissertation**. I am thankful for their aspiring guidance, invaluably constructive criticism and friendly advice during the work. I am sincerely grateful to them for sharing their truthful and illuminating views on a number of issues related to the research.

DECLARATION

I hereby declare that the Dissertation proposal entitled "**Identifying effective attributes for structuring social groups in Social media using Principal Component Analysis"** submitted for the M.Tech Degree is entirely my original work and all ideas and references have been duly acknowledged. It does not contain any work for the award of any other degree or diploma.

Date: Investigator: **Samiya Shafi**

Registration No: 11307674

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

This chapter gives an introduction about Big Data, Social media data and also discusses about the Principal component Analysis technique. The Principal Component Analysis used in identifying effective attribute in making social circle have been discussed and this technique have been illustrated with examples.

1.1 Introduction to Big Data:

Big Data refers to large, complex and growing volume of data sets with evolving relationships from heterogeneous and autonomous sources. Over 2.5 quintillion bytes of data are created everyday and 90% of the present data has been created in the last two years [Wang, 2004].

Big Data is generated from number of data intensive application like online discussions, Flicker (public picture sharing site),sensors, online shopping sites, social media giants(facebook, twitter ,LinkedIn ,YouTube ,Google+ and more),scientific data analysis, mobile devices and more [Laurila,2012]. Each day Google has 1 billion above queries, Twitter has 250 million above tweets, Facebook has 800 million above updates, and YouTube has more than 4 billion views per day [Thakur, 2014]. Nowadays data is produced is in zettabytes (10^{21} bytes).

Data collection has grown tremendously and traditional software's are not able to capture, manage and process within elapsed time. Data becomes big data when traditional software's were unable to ingest, store, analyze and process large volume of data and diverse variety of data or its fast velocity.

Big Data is collection of various forms of data which can be classified as structured data and unstructured data which are defined below:

- **Structured Data:** Structured data is stored in fixed field within a record or file. Structured data contains data from traditional relational database and spreadsheets. For structured data model is created that defines what fields of data will be stored and how data will be stored.

- **Unstructured Data:** Unstructured data can't be easily classified, interpreted and does not fit into traditional databases. Unstructured data are photos, videos, webpage's, pdf files,

  emails, blogs entries, wikis ,word processing documents and PowerPoint presentations, and social media posts. Big Data consists of 85% unstructured data.

Big Data analysis is to find patterns, relationships, predictions, gain business insight and deliver actionable intelligence from large volume of data.

1.1.1 Characteristics of Big Data

The characteristics of big data can be defined under five V's. These provide the various features of Big Data. The V's can be expanded as:

- Volume
- Velocity
- Variety
- Value
- Veracity

i) Volume

Increase in volume of data is due to many factors which can be listed as:

- Machine generates volume of data and amount of data generated by machines are more than non-traditional data.
- Transaction-based data stored from past years.
- Unstructured data from social media.
- Increasing amount of data collected by sensors and machine-to-machine.
- Data generated by single jet engine in 30minutes is 10TB. For 25,000 airline flights per day will generate data in petabytes.

ii) Velocity

Velocity is how fast data is generated and processed. Data is streaming at fast speed must be dealt in a timely manner. Dealing with the velocity of data generated is another challenge. At 140 characters per tweet, the high velocity of Twitter data ensures large volumes 8 TB per

day. Social media data streams produce valuable relationship and opinions for improving customer relationship management.

iii) Variety

Traditional data sources produce data that have well defined fixed data formats .Nontraditional data sources like sensors, social media and more, produce data in all formats structured, unstructured and semi-structured. New data types are needed to store information from non traditional sources. Unstructured data is stored in NoSqL databases.

iv) Veracity

The collected data's qualities vary. Veracity refers to biases, noise and abnormality in data.

Accuracy of analysis depends on the veracity of the source data. Is the data that is stored and mined related to problem being analyzed. Keep data clean and processes to keep 'dirty data ' for accumulating.

v) Value

Different data has different economic value. Meaningful value is hidden among volume of data, the challenge is to find what is valuable and then transform and extract data for analysis [Derakhshan,2007].

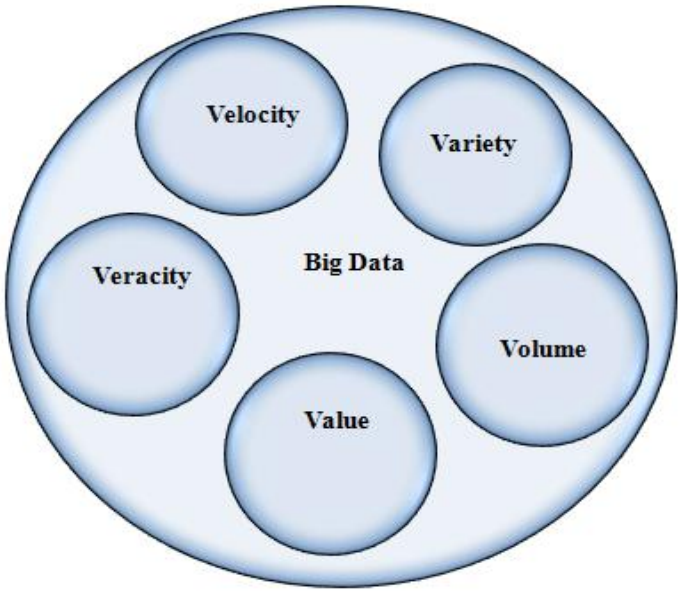

Figure 1.1: V's of Big Data

1.1.2 Challenges in Big Data

Big Data starts with large-volume of data collected from heterogeneous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data [Wu,2014].

i) Heterogeneous and diverse sources:-

Data is collected from different sources and each source have their own schemata or protocols, data representation. Challenges faced in heterogeneous source are different types of representation for same item.

ii) Autonomous sources with distributed and decentralized control:-

Each data source independently generates and collects information without any centralized control.

iii) Complex and evolving relationships:-

In dynamic world relationship between individuals and social ties may evolve with respect to temporal, spatial, and other factors.

iv) Storage challenges

Distributed File system is possible solution for storing volume of data. Distributed System will provide scalability and push computation on various nodes in cluster. Storage challenges are due to variety (structured and unstructured) data. Solution to this challenge is NoSqL databases. NoSqL stores data of all formats.

v) Incompleteness

In data analysis incompleteness of data must be managed. Consider a hospital database design that has fields for patient name, age, occupation, medicine prescribed, disease and blood type for each patient. What if one or more field value is not provided by patient? Still patients data will be stored in database, but missing values will be set to NULL. A data analysis that classify patients according to their occupation, must take into account patients whose attribute value are missing [Labrinidis,2012].

vi) Scale

Managing increasing volume of data has been challenging issue from past many decades. In past challenge was reduced by increasing processors speed. But now data volume is increasing faster than computing resources and processing speed. New large data processing

systems deal with parallelism on single node. Scalability limitation can be removed by cloud computing, which aggregates multiple resources with different workload and varying performance goals. Old HDDs are continuously replaced by new storage technologies such as solid state drives and Phase change Memory.

vii) Timeliness

The speed of processing dataset is inversely proportional to volume of dataset. The larger dataset for processing will take longer time to analyze. In many real time situations result of analysis is required fast. Such as in fraudulent credit card transactions, the suspect should be flagged before transaction is completed. Another example is traffic management system which contains information about number of vehicles and hot spots on roadways. System predicts road congestion points on the route chosen by user and suggests alternate routes. System evaluates multiples spatial proximity queries to find congestion and for suggesting alternate routes to users. Designing systems with fast querying response, when data volume is growing is challenging.

viii) Privacy

Strict laws govern privacy of information, what can be done and what cannot be done on data. Public have great fear regarding misuse of their personal data. Managing privacy is both technical and sociological issue .Privacy issue must be resolved together from both perspective .An attacker can use information publicly available for attacks. Another challenge is user do know what how to share private data with limited disclosure of data.

ix) Human Collaboration

In spite of advanced computational analysis, still many patterns are only recognized by humans easily but computer algorithms find it hard to recognize those patterns. Big Data analysis system take input from multiple human experts.

A popular new method of harnessing human ingenuity to solve problems is through crowd-sourcing. The input provided by humans is mostly correct , but human can provide false information also to mislead .We need technology to identify whether human has enter correct or wrong information.

1.2 Big Data Analysis :

Data analysis is process of knowledge discovery from data. Cleaned data should be processed for analysis to output good valuable result [Sanjay**,**2013].

1.2.1 Steps for Big Data analysis

i) Data Collection

All data that is to be processed is collected for analysis. Difficulties in data collection are different formats of data, collected from heterogeneous sources. After data collection data integration is performed.

ii) Data Cleansing

Data collected may be erroneous, noisy or may contain missing values. Data cleaning uses different methods to remove bad data from dataset. After data cleaning, data may be preceded for data analysis.

iii) Data Analysis

In data Analysis different analytic methods and techniques are used. These methods and techniques can be divided into statistical analysis, data mining and machine learning. Data Analysis is used to discover hidden relationships, patterns, predicts models that are present within the data.

1.3 Social Media Data:

Social networking sites have hundreds of millions of users. Social networks are tool for connecting people, and mirror real-life relationships and society. Users of social media maintain profile information like location, education; concentration, birthday, education; class; school, first_name, hometown, and much more[Ellison,2007]. Cost and overhead previously make communication unfeasible, but advances in social networking technology have made sharing possible [Bae,2004].

1.3.1 Overview of social media data

Social media enable users to communicate information to anybody who is looking for it. The job market is also looking at the information people share about themselves on sites like LinkedIn and Facebook. Examples of mature social data are Twitter and Facebook. On sending a message on social media is as simple as sending an SMS text message. Twitter users sends tweet can be read by the entire world. Twitter focuses on C2W (customer to world) communication. Facebook focuses on interactions between friends, C2C (customer to customer). There is a massive amount of content being created and shared continuously across multiple networks and in various formats. Social media sites like Google+, Facebook, Twitter and so on, have large number of users.

1.3.2 Social media giants

There are many social media sites, which generate large among of data every day such as **Google+ ,YouTube, Facebook, Twitter.**

i) Google+: Google+ introduces Circles, Hangouts and many other unique features. Google+ has 540 million monthly active users [Goga,2014] and sharing 1.5 billion photos each week. Google+ is fifth social networking site in world.

ii) YouTube: YouTube is the world's famous video-sharing site[Cheng,2008]. YouTube is available in 61 languages versions through user interface [Chin, 2007]. More than 1 billion unique users visit YouTube each month.

iii) Facebook: Facebook helps its members to connect and share with the people. It allows users to like, share, videos chatting and comment on pictures, videos, website links, articles and more[Ko,2010]. Facebookcurrently has over 650 million active users every day.

Most social network sites provide an aggregated stream of social news about all of one's ties; this is known as the "News Feed" on Facebook. The feed contains a frequently updated stream of ties' recent activity.

iv) Twitter: Over half a billion tweets are generated per day. Twitter is world's go-to source for connecting about recent updates.

Social Network is powerful means of sharing, organizing and finding contents, contacts[Mislove,2007].Users in social networking sites join network, create profiles and relationships with any users of same social network with whom they associate. Social networks revolve around users, user group friends into circle in Google+ and list in Facebook.

Social networking sites allow users to categorize their friends manually into their social groups, circles and in their social list. Manually categorization of friends in lists and circles is time consuming and lengthy [Mcauley,2012].Circles can be used for content filtering, privacy and sharing data between its users [Smith,2014]. Data is collected from social networking sites about user through different attributes like "tag", "comments", "like", "status", "photos" and "video" which is refer as social media data. These data are the basis for creating models of the relationships between users. They can be used to significantly increase the relevance of what is shown to the user, for advertising and marketing purposes of products [Hochreiter,2014].

World Wide Web, Social Web and mobile devices generate location-based and context-specific content [Alt,2014].Social data is also relevant for companies: complaints, suggestions, opinions. Social networking sites are popular. They provide opportunity to understand characteristics of online social network graph at large scale .Understanding social network graph help to improve current system and design new applications.

1.3.3 Tools and technique for social data analysis are

I. **Social Media Monitoring:** Social Media Monitoring is Social media intelligence tools for collecting sentiment information, frequently discussed topics and more.
II. **Text mining:** Text mining is a promising discipline for Social Media analysis tools for unstructured data. Text mining encompasses the preprocessing of documents, techniques to analyze these intermediate representations as well as the visualization of the results. Text mining contains vast algorithms and techniques.

1.4 Principal Component Analysis:

Principal Component Analysis is a procedure that transforms number of correlated variables into smaller number of uncorrelated variables called principal components [Moore,1981].The goal is to reduce number of variables but retain original meaning of the data[Shlens,2014],[Kolenikov,2013]. PCA is equivalent in finding direction axis which have maximum variance, then using these new directions to define new basis.

Let us consider a multivariate dataset that is represented in terms of a $n \times m$ matrix, $A_{n,m}$ where m columns are sample and n rows are variables.

$$A_{n,m} = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,m} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,m} \end{bmatrix}$$

Then matrix A is linearly transform into another matrix B of same dimension $n \times m$,so that for some $n \times n$ matrix, C.

$$B = CA \qquad \text{Eq. 1}$$

The above equation represents change of basis. Let us consider the row of C to be row vectors c_1, c_2, $c_3 \ldots c_n$ and column of A to be column vectors a_1, a_2... a_m then Eq. 1 can be represented as

$$B = CA = \left(Ca_1, Ca_2, Ca_3, \ldots Ca_n\right) = \begin{bmatrix} c_1a_1 & c_1a_2 & \cdots & c_1a_m \\ c_2a_1 & c_2a_2 & \cdots & c_2a_m \\ \vdots & \vdots & \ddots & \vdots \\ c_na_1 & c_na_2 & \cdots & c_na_m \end{bmatrix}$$

The $c_i a_j \subset R^n$ and $c_i a_j$ is Euclidean dot product. The $c_i a_j$ representation means that original data, A is projected on to rows of C. Now, the columns of C that is (c_1, c_2, $c_3 \ldots c_n$) are new basis for showing rows of A. The columns of C then are principal component directions.

In order to represent the independence between principal components in new basis the variance of the data in the original basis is considered. Original data is de-correlate by finding the direction in which variance is maximized .These direction are used to define the new basis as defined by new basis C.

The variance of random variable, W with mean μ is given by following equation

$$\sigma_W^2 = \mathrm{E}\left[(W-\mu)^2\right] \quad \text{Eq. 2}$$

Where W=random variable, μ=mean, σ_W =variance of W random variable. Considering a vector $\tilde{u}$ of n discrete measurements, $\tilde{u} = \tilde{u}_1, \tilde{u}_2, \tilde{u}_3 ....., \tilde{u}_n$ with mean of u vector is μ_u .A translated set of measurements $u = u_1, u_2, u_3 ....., u_n$ can be obtained by subtracted mean from each measurement that has zero mean. The variance of these the measurement u can be given by following equation

$$\sigma_u^2 = \frac{1}{n-1}\left(uu^T\right) \quad \text{Eq. 3}$$

If second vector $v = (v_1, v_2, v_3 .... v_n)$ of n measurements again with zero mean is considered then it can be generalized to obtain covariance by considering both u and v as given in Eq. 4. *A c*ovariance can be used to measure how much two variables change together, covariance can be positive or negative. In PCA covariance is positive covariance, close to zero. Variance is special case of covariance, when variables are similar. In PCA variance are maximized and covariance are minimized. The variance between u and v vector is as follows:

$$\sigma_{uv}^2 = \frac{1}{n-1}\left(uv^T\right) \quad \text{Eq. 4}$$

Generalize the matrix $A_{n,m}$ can be used to represent data matrix.

$$A_{n,m} = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,m} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,m} \end{bmatrix} = [a_1 \quad a_2 \quad \cdots \quad a_m] \in R^{n\times m}, a_i^T \in R^m \quad \text{Eq. 5}$$

Column vector for each variable contain samples for one particular variable. Here a_i is vector for ith variable of *n* samples. Covariance matrix,$m \times m$ can be computed using following equation

$$C_A = \frac{1}{n-1} AA^T = \frac{1}{n-1} \begin{bmatrix} a_1 a_1^T & a_1 a_2^T & \cdots & a_1 a_m^T \\ a_2 a_1^T & a_2 a_2^T & \cdots & a_2 a_m^T \\ \vdots & \vdots & \ddots & \vdots \\ a_m a_1^T & a_m a_2^T & \cdots & a_m a_m^T \end{bmatrix} \in R^{m\times m} \quad \text{Eq. 6}$$

Where C_A is covariance. The covariance between m variable is computed using the above equation. Covariance matrix is symmetric and square matrix. Compute all covariance pairs between m variables. Off-diagonal values are covariances between m pairs and principal diagonal values are variances.

The dataset can be linearly transformed , A into B using equation B=CA. Make supposition

function transformed matrix, B to exhibit and relate supposition features to covariance matrix C_B .Supposition is that transformed matrix, B have uncorrelated variables ,that is matrix C_B have covariance of variables as possible close to zero. Requirement for covariance matrix C_B are to maximize variance and to minimize off-diagonal covariance between variables. Result from this supposition is that find covariance close to zero, C_B .Choose transformation matrix, C with principal diagonal C_B ,then objectives of PCA are achieved.

Consider new supposition that vector $c_1, c_2, c_3, \ldots . c_n$ are orthogonal. The covariance matrix C_B and B can be represented as follows

$$C_B = \frac{1}{n-1} BB^T = \frac{1}{n-1}(CA)(CA)^T = \frac{1}{n-1}(CA)(C^T A^T) = \frac{1}{n-1} C(AA^T)C^T$$

$$C_B = \frac{1}{n-1} CXC^T \qquad \text{Eq. 7}$$

where $X = AA^T$, X is $m \times m$ symmetric matrix, $(AA^T)^T = (A^T)^T (A)^T = AA^T$ .Theorem of linear algebra is applied which states that *every square symmetric matrix is orthogonally diagonalizable*. Theorem can be represented as

$$\mathrm{X} = GPG^T \qquad \text{Eq. 8}$$

Where G is $m \times m$ orthogonal matrix whose columns are orthogonal eigenvectors of X, and principal diagonal, P entries are eigenvalues of X. The highest eigen value is first principal component (PC1), the second highest is second principal component (PC2), and so on [Lukibisi,2010].

After computing eigenvalues and eigenvectors of $X = AA^T$, sort eigenvalues in descending order and place these eigenvalues on diagonal entries, P. Now construct an orthogonal

matrix, G by placing eigen vectors with highest eigen value in first column, the eigenvector of second highest in second column and so on[Comon,1994][Abdi,2010].The objective of diagonalising covariance matrix, of transformed data is achieved. The principal components (columns of C) are Eigen vectors of covariance matrix,AA^T and columns of P are in decreasing order of 'importance'. The first column is more meaning full as compared to other columns.

Chapter Organization:

Chapter 1: The chapter gives a brief introduction about Big Data, social media data and Principal Component Analysis.

Chapter 2: The chapter provides a detailed literature survey and work done in the field of Big Data.

Chapter 3: The chapter gives an insight of the problem definition of research work, its objectives and methodology.

Chapter 4: The chapter gives the description of the various work scenarios to carry out the research work.

Chapter 5: The chapter gives the results obtained at every stage of analysis.

Chapter 6: The chapter gives a brief discussion about the conclusion derived from research work and its probable future scope.

CHAPTER 2
LITERATURE SURVEY

This chapter is a brief discussion about the various pioneering works in the field of Big Data and the various people who made contributions in similar field. From its origin to various developments in analysis, this chapter also contains the techniques used in analysis.

Alan Mislove et at.(2007) purposes large-scale measurement study and analysis of online social networks. This model obtains social network graph through crawling user links. Result shows power law, small world ,and scale properties of online social networks. When indegree of nodes match its outdegree then that network is densely connected. Structure of online social network can be found by focusing on weakly connected components(WCC). Forward and Reverse links help us to crawl whole WCC. Forward links does not crawl entire WCC. Both forward and reverse links crawl entire large WCC. The result shows that level of symmetry in Flicker, LiveJournal and YouTube with directed links have significant degree of symmetry. Analysis shows that degree distributions in social media networks differs from power-law network. Analysis show that low degree nodes are part of few groups and high-degree nodes are part of multiple groups.

Lukibisi et at.(2007) analyzes mineral composition data using principal component analysis(PCA). This paper presents when, how and why PCA works. Dataset contains number of dimensions or variables. Superfluous variables increases data collection and cost for data processing of deploying model. Principal Component Analysis is mathematical procedure to reduce number of dimensions of dataset but does not lose its original variability in data. Result shows that scree test plotted Eigen values associated with component. Break will appear between components with relatively large eigenvalues and components with small eigenvalues. The component that appear before break are consider to be valuable and are retained for rotation and those who appear after the break are consider unimportant and are not retained. Dimensions which have greater amounts are major parts of component.

Shanli Wang et at.(2009) presents structure model on data mining based on neural network in detail , key technology and ways to achieve data mining based on neural networks. Combination of data mining method and neural network model improve efficiency of data mining. Data mining based on neural network is divided into phases: model option, data preparing, rules option and result assessment. There are number of types of data mining based on neural network. Mostly used type is data mining on fuzzy neural network .Fuzzy neural network model structure has different layers. The input layer is for inputting data into model, then affiliation relation is computed on inputted data, then data preprocessing is done on data, data preprocessing is to enhance cleaned data which has been selected. Rule option, there are many methods of extracting rules such as extract fuzzy rules method and black-box method. At last result is computed and outputted.

Jiong Xie et at.(2010) analyzed that ignoring data locality in heterogeneous computing nodes in a cluster can reduce MapReduce performance. The proposed model reduces data movement between nodes to improve performance of Hadoop in heterogeneous clusters, aim can be achieved by data placement scheme. In Data placement management algorithms are implemented on HDFS. Firstly algorithm distributes file fragments to nodes in cluster according to node's computing ratios, then other algorithm is for reorganizing file fragments. Data distribution server creates over-utilized node list and underutilized node list. Data distribution server continuously transfer file fragments from over utilized to underutilized nodes to balance load on nodes. Result shows that response time of grep and wordcount on nodes in heterogeneous cluster is less for modeled data placement strategy as compared other data placement strategy.

Danah boyd et at.(2011) discussed some scenario which provokes question for cultural ,technological ,and scholarly phenomenon for Big Data. With automatic research, the definition of knowledge has changed. The volume of data has increases and new methods and techniques find knowledge. Big Data is changing objects of knowledge. Massive amount of data and applied mathematics replace old tools. Special tools for big data also have limitations and restrictions. Claims made by other disciplines mislead goal and accuracy of knowledge. Large dataset from different sources are unreliable and when data are integrated from different sources, errors get magnified. Bigger data are not always better data. Data are

not equivalent, they cannot be analyzed in similar manner. Public accessible data doesn't make it ethical. Compromising privacy from public accessible data is unethical .Accessing only particular part of data creates new digital divides.

Oracle (2013) This paper is about Oracle Advanced Analytics. Oracle presents integrated architecture for big data analysis that makes task easier to perform and reduce data movement among integrated components. Analyzing social media data shows speed at which sentiments can shift online. Pre-processing social media data with Hadoop predicts customer behavior, anticipate cross/up-sell opportunities, improve marketing and more. When data is loaded in Oracle database, Oracle Advanced Analytics uncover hidden relationships in data. Oracle Advanced Analytics is powerful combination of in-database algorithms and R algorithms. Oracle database have capabilities to ensuring security, scalability efficiency, reduce data movement. Oracle Advanced Analytics when used with Hadoop and MapReduce delivers data to acquire, organize, analyze and maximize value of big data with least data movement and high security.

Sanjay P.Ahuja et at.(2013) paper presents current trends and characteristics of Big Data, its analysis and challenges in data collection, storage and management in cloud computing. Hadoop processes large amount of data, using batch processing approach. Pig and Hive provide simplifying querying. NoSqL manages big data, by storing data unstructured data. Big Data is new way for finding interesting patterns, meaningful information. Facebook, Twitter and LinkedIn use Hadoop for processing large amount of data.

Alt et at.(2014) suggests an approach which increases efficiency of defining ontologies by automatically discovering knowledge from existing business databases. With ontology engineering approach ontologies are created and combined with text mining tool .Text mining mines the data from different document-types such as emails, new articles and consumer reviews, provides key procedure to interpret text. Purposed framework of an ontology-based social media analysis contain business databases, ontology builder, text mining component and social media. Result of analysis is stored for sentiment analysis and for complaints or positive feedback. Ontology Transformation algorithms are DB2OWL, R_2O, RDB2ONT and RONTO. DB2OWL automatically map relational database schema, unstructured and structure data to ontologies. Different algorithms are available for according

to requirement of ontology engineering process. There is no algorithm in present for transformation of object-oriented databases.

Julian McAuley et al. (2014) proposed model for identifying users' social circle using unsupervised machine learning algorithm. Parameter vector encodes what similar attributes of profile form circle. Profile information is encoded as tree. Modeling users membership to multiple groups, identifies overlapping and hierarchically nested groups. Evaluating difference vector based on parents of leaf node will result what profile categories two users have common. Based on difference vectors, describe how to construct edge features.
Four different ways of representing similarity between different profiles for two users was also discussed. The ways for evaluating difference vector and identifying compatibility between pair of profiles. Alignment between predicted circle and ground circle is measured by Balanced Error Rate, optimal match through linear assignment. Result found that all algorithm work better on Facebook than on Google+ or Twitter.

Xindong Wu et at.(2014) proposed the Big Data processing model using data mining and HACE theorem that characterizes features of Big Data. The framework at conceptual view is divided into different tiers. Firstly the data is accessed and computed and then the data privacy is discussed through privacy-preservation approaches and encryption mechanism, and then Big Data algorithms on multisource are discussed such as theory of local pattern analysis. Big Data processing model depends on MapReduce and cloud computing platform. Data mining algorithm is integrated with MapReduce to improve processing and scalability. Privacy Protection in Big data summarizes methods for privacy include summarization, suppression and data swapping. Developing safe sharing protocol in Big Data is challenging. Big Data is emerging technology need system designed that link data, their relationships to predict patterns and hold large volume of data.

CHAPTER 3

PRESENT WORK

This chapter gives an insight of the objectives of work which will be carried out in identifying attributes in social media data and identifying technique used for the same.

3.1 Proposed Work:

In social media, user attributes is used as profile information which help user to form social circle. User add friends into circles and list by categorizing them using profile information and making it easy to filter the friends. The user in circle has some particular number of attributes, which help them to make circle. The study proposes an approach on finding attributes which effectively contribute in making the circle using Principal Component Analysis.

Let us consider *n* as numbers of attributes which define a user in social media for creating circles. Principal Component Analysis (PCA) helps to identify attributes which are meaningful and reduce dimensionality of data. With PCA the complexity of data can be reduced, need less number of plots to analyze. Principal Component Analysis is mathematical procedure to reduce number of dimensions (attributes) of dataset but maintaining its original variability. Interpretation and validation of the proposed methodology will be plotted by scree test.

Instead of working on all m-dimensions, first perform PCA on original data (m-dimensions) then use only first few components say PC1, PC2 in analysis. These reduced dimensions can be used to focus on making circle. Circle is made with users, users have number of attributes focusing on all the attributes is very lengthy, costly. With PCA reduced attributes, analysis can be easy, and focusing on only those attributes which are efficient in making circles.

3.2 Objectives:

The objectives of this dissertation are as follows:

1. Study and Analysis of various approaches for identifying social circles or groups in Social media data.

2. Examine the Social media analysis model to find which attributes contribute in making circles on Principal Component Analysis.
3. To model and propose an appropriate and efficient method using Principal Component Analysis for Social media Analysis.
4. To implement the proposed model using R and validate its accuracy using scree test.

3.3 Research Methodology:

Research is a logical, methodical and orderly search for new and useful information on a specific subject. It adds contribution to the existing knowledge. Research Methodology is one of the ways to interpret and resolve a research problem.

In order to achieve the mentioned objectives, the effective methodologies are considered to complete this task as:

1. In order to achieve the objective **"Study and Analysis of various approaches for identifying social circles or groups in Social media data"**, comprehensive literature survey was carried out for Big Data Analysis techniques, different challenges were observed that exist in big data and privacy issues.

2. In order to achieve the objective **"Examine the Social media analysis model to find which attributes major contribute in making circles in social media by using Principal Component Analysis"**, the work is lead as:

a) First, a complete analysis of the social media circles is done.

b) Then with Principal Component Analysis approach find which attributes contribute in making circles.

c) Principal Component Analysis approach was critically reviewed.

3. In order to achieve the objective **"To implement the proposed model using R and validate its accuracy using scree test"** model is designed using Principal Component Analysis and scree test.

a) Division of dataset: Select the dataset required for finding social data network.

b) Design: Perform cleansing on social data, apply Principal Component Analysis on cleansed data.

c) Apply scree test to plot attributes which significantly effect in identifying groups.

The flowchart of research methodology is given in figure 3.1.

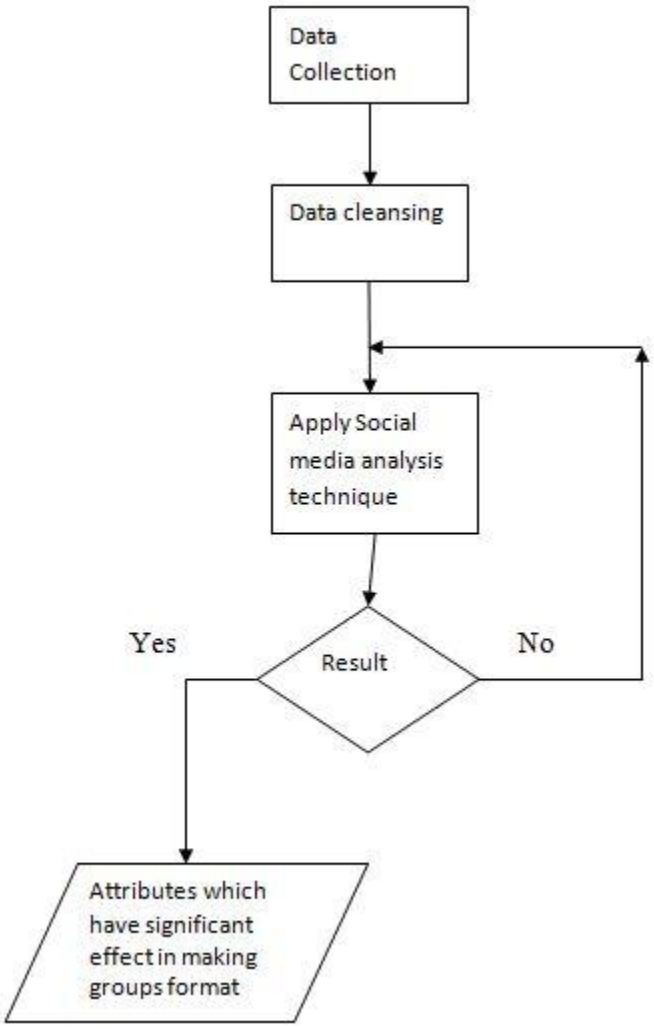

Figure 3.1: Flow chart of research methodology

CHAPTER 4

EXPERIMENTAL SETUP

This chapter lists the various steps proposed for identifying effective attribute in structuring social circles in social media and a work timeline for same. Implementation of PCA is done in R programming. R programming is environment and programming language for graphics and statistic computing. It is used by data miners, data analysts and statisticians. R implements number of techniques like classification, clustering, statistics techniques. R is extensible through installation of add-on packages.

4.1 Identification Phases:

Principal component analysis is distribution of variation of multivariate dataset across components in such a way to find patterns that may not observe using other analysis and graphics technique. It is a procedure that transforms number of correlated variables into smaller number of uncorrelated variables called principal components [Moore,1981].The goal is to reduce number of variables but retain original meaning of the data.

The entire process of identifying effective attribute in structuring social circles in social media can be represented as following steps [Lukibisi et at,2007].

Step 1: Preparing Data

The initial phase of analysis of social media data is collection of data. The data for analysis is collected from profile information of users and their circles. In Social media data, profile information of user has many dimensions such as name, school, place, location, birthday, degree, class, school, name, hometown etc.

Consider the dataset with 244 attributes which are used for building the profile of the user in Google+ and facebook [Leskovec,2014]. The dataset consists of number of circles, their users and attributes of the users. Attribute value are either '1' or '0', '1' means attribute consists of a value,'0' means attribute does not contain any value. The attribute values are anonymized for privacy concerns. Profile information of a user is represented in tree structured as given in figure 4.1 which consists of the attributes as described in the dataset.

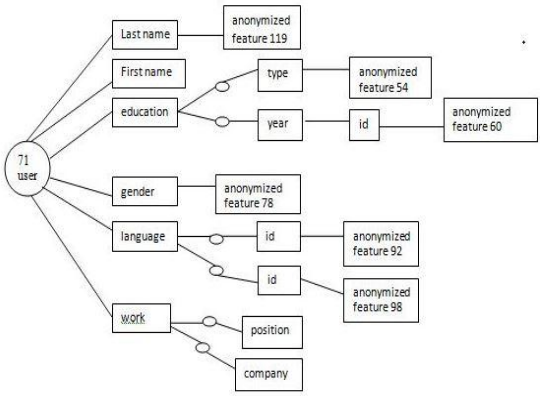

Figure 4.1: User71 attribute tree

Table 4.1: Attribute Present in user71 and attribute name

AttributePresent	**AttributeName**
1	Lastname:anonymizedfeature119
0	Firstname
1	Education:type:anonymizedfeature54
1	Education:year:id:anonymizedfeature60
1	Gender:anonymizedfeature78
1	Language:id:anonymizedfeature92
1	Language:id:anonymizedfeature98
0	Work: position
0	Work: company

Social media dataset for circle0 is represented in terms of a $n \times m$ matrix, $A_{n,m}$ where m columns are attributes and n rows are users.

$$A_{n,m} = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,m} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,m} \end{bmatrix}$$

Then matrix A is linearly transform into another matrix B of same dimension $n \times m$,so that for some $n \times n$ matrix, C.

$$B = CA \qquad \text{Eq. 1}$$

The above equation represents change of basis.

The variance of random variable, W with mean μ is given by following equation

$$\sigma_W^2 = \mathrm{E}\left[(W-\mu)^2\right] \quad \text{Eq. 2}$$

Where W=random variable, μ=mean, σ_W = variance of W random variable.

Step 1.1: Cleaning of data

The attributes which have zero users are removed from dataset. After cleaning dataset dimension will be reduced.

Step 1.2: Scaling of data

Scaling of data is to standardize and centralize the data. Scaled data can be obtained by subtracted mean from each measurement, which will result data with zero mean. Considering a vector $\tilde{u}$ of n discrete measurements, $\tilde{u} = \tilde{u}_1, \tilde{u}_2, \tilde{u}_3 ....., \tilde{u}_n$ with mean of u vector is μ_u .A translated set of measurements $u = u_1, u_2, u_3 ....., u_n$ can be obtained by subtracted mean from each measurement that has zero mean. The variance of these the measurement u can be given by following equation

$$\sigma_u^2 = \frac{1}{n-1}\left(uu^T\right) \quad \text{Eq. 3}$$

If second vector $v = (v_1, v_2, v_3 .... v_n)$ of n measurements again with zero mean is considered then it can be generalized to obtain covariance by considering both u and v as given in Eq. 4.

Step 2: Calculate the covariance matrix between users of circle and attributes of profile information

*A c*ovariance can be used to measure how much two variables change together, covariance can be positive or negative. In PCA covariance is positive covariance, close to zero. Variance is special case of covariance, when variables are similar. In PCA variance are maximized and covariance are minimized. The variance between u and v vector is as follows:

$$\sigma_{uv}^2 = \frac{1}{n-1}\left(uv^T\right) \quad \text{Eq. 4}$$

Generalize the matrix $A_{n,m}$ can be used to represent data matrix.

$$A_{n,m} = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,m} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,m} \end{bmatrix} = [a_1 \quad a_2 \quad \cdots \quad a_m] \in R^{n\times m}, a_i^T \in R^m \qquad \text{Eq. 5}$$

Column vector for each variable contain samples for one particular variable. Here a_i is vector for i[th] variable of n samples. Covariance matrix,$m \times m$ can be computed using following equation

$$C_A = \frac{1}{n-1} AA^T = \frac{1}{n-1} \begin{bmatrix} a_1 a_1^T & a_1 a_2^T & \cdots & a_1 a_m^T \\ a_2 a_1^T & a_2 a_2^T & \cdots & a_2 a_m^T \\ \vdots & \vdots & \ddots & \vdots \\ a_m a_1^T & a_m a_2^T & \cdots & a_m a_m^T \end{bmatrix} \in R^{m\times m} \qquad \text{Eq. 6}$$

Where C_A is covariance. The covariance between m variable is computed using the above equation. Covariance matrix is symmetric and square matrix. Compute all covariance pairs between m variables. Off-diagonal values are covariances between m pairs and principal diagonal values are variances.

The dataset can be linearly transformed , A into B using equation B=CA. Make supposition

function transformed matrix, B to exhibit and relate supposition features to covariance matrix C_B .Supposition is that transformed matrix, B have uncorrelated variables ,that is matrix C_B have covariance of variables as possible close to zero. Requirement for covariance matrix C_B are to maximize variance and to minimize off-diagonal covariance between variables. Result from this supposition is that find covariance close to zero, C_B .Choose transformation matrix, C with principal diagonal C_B ,then objectives of PCA are achieved.

Step 3: Compute Eigen Values and Eigen vector from the covariance matrix

Consider new supposition that vector $c_1, c_2, c_3, \ldots . c_n$ are orthogonal. The covariance matrix C_B and B can be represented as follows

$$C_B = \frac{1}{n-1} BB^T = \frac{1}{n-1} (CA)(CA)^T = \frac{1}{n-1} (CA)(C^T A^T) = \frac{1}{n-1} C(AA^T)C^T$$

$$C_B = \frac{1}{n-1} CXC^T$$

Eq. 7

where $X = AA^T$, X is $m \times m$ symmetric matrix, $(AA^T)^T = (A^T)^T (A)^T = AA^T$ .Theorem of linear algebra is applied which states that *every square symmetric matrix is orthogonally diagonalizable*. Theorem can be represented as

$$X = GPG^T$$

Eq. 8

Where G is $m \times m$ orthogonal matrix whose columns are orthogonal eigenvectors of X, and principal diagonal, P entries are eigenvalues of X. The highest eigen value is first principal component (PC1), the second highest is second principal component (PC2), and so on [Lukibisi,2010].

Step 4: Compute Principal Components

After computing eigenvalues and eigenvectors of $X = AA^T$, sort eigenvalues in descending order and place these eigenvalues on diagonal entries, P. Now construct an orthogonal matrix, G by placing eigen vectors with highest eigen value in first column, the eigenvector of second highest in second column and so on[Comon,1994][Abdi,2010].The objective of diagonalising covariance matrix, of transformed data is achieved. The principal components (columns of C) are Eigen vectors of covariance matrix,AA^T and columns of P are in decreasing order of 'importance'. The first column is more meaning full as compared to other columns.

Step 5: Result Interpretation

The results of PCA are plotted in R programming. The Plots shows the attributes for which users of circle0 are highly correlated.

CHAPTER 5

RESULTS AND FINDINGS

This chapter includes all the results obtained during each and every phase of Identifying effective attributes for structuring social groups in Social media. This chapter is concluded with the issues faced during the process and future scope of this research.

5.1 Identifying Attribute Phases Results:

The step wise results of the various identifying attribute phases are as follows:

Step 1: Preparing Data

The dataset consists of 24 circles. Every circle is formed from at least 1 user and ranging till 133 users as given in table 5.1. Let us take an example from dataset **circle1** contains only **one user** and **circle15** contains **133 users**. The same user can be in different circles or in one circle. **User258** in dataset is present in **circle4** and **circle16**.Circles tend to overlap is they comprise of users with similar id. As in the dataset circle 1 and 16, 8 and 20 overlap each other as shown in table 5.2.

Table 5.1: Circle names with their user count

Circle name	**Circle 0**	**Circle 1**	**Circle 2**	**Circle 3**	**Circle 4**	**Circle 5**	**Circle 6**	**Circle 7**	**Circle 8**	**Circle 9**	**Circle 10**	**Circle 11**
No. of users in circle	20	1	9	3	17	1	20	2	1	10	4	30
Circle name	**Circle 12**	**Circle 13**	**Circle 14**	**Circle 15**	**Circle 16**	**Circle 17**	**Circle 18**	**Circle 19**	**Circle 20**	**Circle 21**	**Circle 22**	**Circle 23**
No. of users in circle	1	5	2	133	32	9	1	13	6	1	1	3

Table 5.2: Circle with their users

Circle name	**User ids of circles**
Circle 4	125,344,295,257,55,122,223,59,268,280,84,156,***258***,236,250,239
Circle 16	251,94,330,5,34,299,254,24,180,194,281,101,266,135,197,***173***,36,9,85,57,37, ***258***,309,80,139, 202,187,249,58,127,48,92
Circle 1	***173***
Circle 8	***282***
Circle 20	244,***282***,262,293,220,174

The **circle8** present in dataset contains only one user that is **user282** and another **circle20** has many users, one user among **circle20** is **user282** as shown in table 5.2. The users which are not connected in form of friends or having a common attribute cannot form the circle. The user should be connected to each other directly or indirectly. The study shows PCA reduce the dimensionality of social data.

The dataset consists of 224 attributes for each user, making it difficult to work on all 224 attributes in formation of circle. In order to analyze the data PCA provides a way to find meaningful attributes which have effective contribution in making circles.

Considering the above dataset, Let us take circle0 for our analysis. Circle0 consists of 20 users with the user ids as **user71, user215, user54, user61, user298, user229, user81, user253, user193, user97, user264, user29, user132, user110, user163, user259, user183, user334, user245, user222**. Implementation of PCA is done in R programming. The users along with their attributes are given in Figure 5.1.

```
> #load dataset of circle0
> library(calibrate)
> originaldataset1<-read.csv(file="originaldata.csv",head=T,sep="\t")
> originaldataset1
   X0attr X1attr X2attr X3attr X4attr X5attr X6attr X7attr X8attr X9attr
1       0      0      0      0      0      0      0      0      0      0
2       0      0      0      0      0      0      0      0      0      0
3       0      0      0      0      0      0      0      0      0      0
4       0      0      0      0      0      0      0      0      0      0
5       0      0      0      0      0      0      0      0      0      0
6       0      0      0      0      0      0      0      0      0      0
7       0      0      0      0      0      0      0      0      0      0
8       0      0      0      0      0      0      0      0      0      0
9       0      0      0      0      0      0      0      0      0      0
10      0      0      0      0      0      0      0      0      0      0
11      0      0      0      0      0      0      0      0      0      0
12      0      0      0      0      0      0      0      0      0      0
13      0      0      0      0      0      0      0      0      0      0
14      0      0      0      0      0      0      0      0      0      0
15      0      0      0      0      0      0      0      0      0      0
16      0      0      0      0      0      0      0      0      0      0
17      0      0      0      0      0      0      0      1      0      0
18      0      0      0      0      0      0      0      0      0      0
19      0      0      0      0      0      0      0      0      0      0
20      0      0      0      0      0      0      0      1      0      0
   X10attr X11attr X12attr X13attr X14attr X15attr X16attr X17attr X18attr
1        0       0       0       0       0       0       0       0       0
2        0       0       0       0       1       0       0       0       0
3        0       0       0       0       0       0       0       0       0
4        0       0       0       0       0       0       0       0       0
5        0       0       0       0       0       0       0       0       0
6        0       0       0       0       0       0       0       0       0
7        0       0       0       0       1       0       0       0       0
8        0       0       0       0       0       0       0       0       0
9        0       0       0       0       0       0       0       0       0
```

Figure 5. 1: Social media dataset

As can be observed, the attributes form columns and users form rows of circle0 dataset. The first row are the attribute values of user71, second row are attribute values of user215 and last 20th row are attribute values of user 222.The dataset contains 224 attributes for each 20 users of circle0. The cleansing of data is removal of those attributes which have zero number of users in circle0. After cleansing 73 attributes of profile information of circle0 remains. Cleaned dataset is given in Figure 5.2.

```
> #load data in R
> dataset1<-read.csv(file="dataset1.csv",head=T,sep="\t")
> dataset1
   X7attr X14attr X23attr X32attr X43attr X50attr X52attr X53attr X54attr
1       0       0       0       0       0       0       0       1       1
2       0       1       0       0       0       0       0       0       1
3       0       0       0       0       0       0       0       0       0
4       0       0       0       0       0       0       0       1       0
5       0       0       0       0       0       1       0       1       0
6       0       0       0       0       0       1       0       1       0
7       0       1       1       0       0       0       1       1       1
8       0       0       0       0       0       0       0       1       0
9       0       0       0       0       0       0       0       0       0
10      0       0       0       0       0       0       0       0       0
11      0       0       0       0       0       0       0       0       0
12      0       0       0       0       0       1       0       1       0
13      0       0       0       0       1       1       0       1       0
14      0       0       0       0       0       0       0       0       0
15      0       0       0       0       0       0       1       0       1
16      0       0       0       0       0       0       0       0       0
17      1       0       0       0       0       0       1       1       0
18      0       0       0       0       0       1       0       1       0
19      0       0       0       1       0       0       0       0       1
20      1       0       0       0       0       1       0       1       0
   X55attr X58attr X59attr X60attr X63attr X65attr X66attr X68attr X77attr
1        0       0       0       1       0       1       0       0       0
2        0       0       0       0       0       0       0       0       0
3        0       0       0       0       0       0       0       0       0
4        0       0       0       0       0       0       0       1       1
5        1       0       1       0       0       1       0       0       0
6        1       0       1       0       0       1       0       0       1
7        1       0       0       0       1       0       0       0       1
8        0       0       0       0       0       0       0       0       1
9        1       0       0       0       0       0       0       0       0
10       0       0       0       0       0       0       0       0       0
```

Figure 5.2: Cleaned dataset

As it can be observed, the number of columns is reduced after cleaning. The Figure 5.3 shows the dimensions of dataset before and after cleansing.

From the Figure 5.3, it is observed that first dimension is number of users in circle and the second dimension is number of attributes .The number of attributes reduces from 224 to 73 after cleaning.

```
> #dimension of dataset before cleaning
> dim(originaldataset)
[1]  20 224
> #dimension of dataset after cleaning
> dim(dataset1)
[1] 20 73
>
```

Figure 5.3: Dimension of original dataset and cleansed dataset

After cleaning **standardize the dataset in R.** The command to standardized dataset in R and scaled data is given in Figure 5.4.

```
> # Scale the cleansed dataset of circle0
> standardize <- function(x) {(x - mean(x))}
> my.stddata = apply(dataset1,2,function(x) (x-mean(x)))
> my.stddata
      X7attr X14attr X23attr X32attr X43attr X50attr X52attr X53attr X54attr X55attr
 [1,]   -0.1    -0.1   -0.05   -0.05   -0.05    -0.3   -0.15    0.45    0.75   -0.45
 [2,]   -0.1     0.9   -0.05   -0.05   -0.05    -0.3   -0.15   -0.55    0.75   -0.45
 [3,]   -0.1    -0.1   -0.05   -0.05   -0.05    -0.3   -0.15   -0.55   -0.25   -0.45
 [4,]   -0.1    -0.1   -0.05   -0.05   -0.05    -0.3   -0.15    0.45   -0.25   -0.45
 [5,]   -0.1    -0.1   -0.05   -0.05   -0.05     0.7   -0.15    0.45   -0.25    0.55
 [6,]   -0.1    -0.1   -0.05   -0.05   -0.05     0.7   -0.15    0.45   -0.25    0.55
 [7,]   -0.1     0.9    0.95   -0.05   -0.05    -0.3    0.85    0.45    0.75    0.55
 [8,]   -0.1    -0.1   -0.05   -0.05   -0.05    -0.3   -0.15    0.45   -0.25   -0.45
 [9,]   -0.1    -0.1   -0.05   -0.05   -0.05    -0.3   -0.15   -0.55   -0.25    0.55
[10,]   -0.1    -0.1   -0.05   -0.05   -0.05    -0.3   -0.15   -0.55   -0.25   -0.45
[11,]   -0.1    -0.1   -0.05   -0.05   -0.05    -0.3   -0.15   -0.55   -0.25   -0.45
[12,]   -0.1    -0.1   -0.05   -0.05   -0.05     0.7   -0.15    0.45   -0.25    0.55
[13,]   -0.1    -0.1   -0.05   -0.05    0.95     0.7   -0.15    0.45   -0.25    0.55
[14,]   -0.1    -0.1   -0.05   -0.05   -0.05    -0.3   -0.15   -0.55   -0.25   -0.45
[15,]   -0.1    -0.1   -0.05   -0.05   -0.05    -0.3    0.85   -0.55    0.75   -0.45
[16,]   -0.1    -0.1   -0.05   -0.05   -0.05    -0.3   -0.15   -0.55   -0.25   -0.45
[17,]    0.9    -0.1   -0.05   -0.05   -0.05    -0.3    0.85    0.45   -0.25    0.55
[18,]   -0.1    -0.1   -0.05   -0.05   -0.05     0.7   -0.15    0.45   -0.25    0.55
[19,]   -0.1    -0.1   -0.05    0.95   -0.05    -0.3   -0.15   -0.55    0.75   -0.45
[20,]    0.9    -0.1   -0.05   -0.05   -0.05     0.7   -0.15    0.45   -0.25    0.55
      X58attr X59attr X60attr X63attr X65attr X66attr X68attr X77attr X78attr
 [1,]   -0.05   -0.15    0.95   -0.15     0.7   -0.05   -0.05   -0.55    0.55
 [2,]   -0.05   -0.15   -0.05   -0.15    -0.3   -0.05   -0.05   -0.55    0.55
 [3,]   -0.05   -0.15   -0.05   -0.15    -0.3   -0.05   -0.05   -0.55    0.55
 [4,]   -0.05   -0.15   -0.05   -0.15    -0.3   -0.05    0.95    0.45   -0.45
 [5,]   -0.05    0.85   -0.05   -0.15     0.7   -0.05   -0.05   -0.55    0.55
 [6,]   -0.05    0.85   -0.05   -0.15     0.7   -0.05   -0.05    0.45   -0.45
 [7,]   -0.05   -0.15   -0.05    0.85    -0.3   -0.05   -0.05    0.45   -0.45
 [8,]   -0.05   -0.15   -0.05   -0.15    -0.3   -0.05   -0.05    0.45   -0.45
 [9,]   -0.05   -0.15   -0.05   -0.15    -0.3   -0.05   -0.05   -0.55    0.55
```

Figure 5.4: Scaled Dataset

```
> mean(my.stddata)
[1] -3.608801e-18
>
```

Figure 5.5: Mean of scaled dataset of circle0

As observed from above figure 5.5, the mean of scaled dataset is approximately near to zero.

Step 2: Calculation of covariance matrix between users of circle and attributes of profile information

As explained in previous chapter, a covariance can be used to measure how much two variables change together, covariance can be positive or negative. Variance is special case of covariance, when variables are similar. The covariance between users of circle and their attributes is given in figure 5.6 and variance of dataset is given in figure 5.7.

The covariance matrix for circle0 can be computed using Equation 6 and variance for dataset can be calculated using Equation 4. Explanation of Equation 4 and Equation 6 is given in previous chapter. The total number of attributes of user profile is 73 after cleansing. So, the covariance matrix for circle0 will be 73x73 matrix.

```
> covariance2=cov(my.stddata)
> # calculate covariance
> covariance2=cov(my.stddata)
> covariance2
              X7attr      X14attr      X23attr      X32attr      X43attr
X7attr     0.094736842 -0.010526316 -0.005263158 -0.005263158 -0.005263158
X14attr   -0.010526316  0.094736842  0.047368421 -0.005263158 -0.005263158
X23attr   -0.005263158  0.047368421  0.050000000 -0.002631579 -0.002631579
X32attr   -0.005263158 -0.005263158 -0.002631579  0.050000000 -0.002631579
X43attr   -0.005263158 -0.005263158 -0.002631579 -0.002631579  0.050000000
X50attr    0.021052632 -0.031578947 -0.015789474 -0.015789474  0.036842105
X52attr    0.036842105  0.036842105  0.044736842 -0.007894737 -0.007894737
X53attr    0.047368421 -0.005263158  0.023684211 -0.028947368  0.023684211
X54attr   -0.026315789  0.078947368  0.039473684  0.039473684 -0.013157895
X55attr    0.057894737  0.005263158  0.028947368 -0.023684211  0.028947368
X58attr   -0.005263158 -0.005263158 -0.002631579 -0.002631579 -0.002631579
X59attr    0.036842105 -0.015789474 -0.007894737 -0.007894737 -0.007894737
X60attr   -0.005263158 -0.005263158 -0.002631579 -0.002631579 -0.002631579
X63attr   -0.015789474  0.036842105  0.044736842  0.044736842 -0.007894737
X65attr    0.073684211 -0.031578947 -0.015789474 -0.015789474 -0.015789474
X66attr   -0.005263158 -0.005263158 -0.002631579 -0.002631579 -0.002631579
X68attr   -0.005263158 -0.005263158 -0.002631579 -0.002631579 -0.002631579
X77attr    0.047368421 -0.005263158  0.023684211  0.023684211 -0.028947368
X78attr   -0.047368421  0.005263158 -0.023684211 -0.023684211  0.028947368
X84attr   -0.005263158 -0.005263158 -0.002631579 -0.002631579 -0.002631579
X90attr    0.042105263 -0.010526316 -0.005263158 -0.005263158 -0.005263158
X91attr   -0.005263158 -0.005263158 -0.002631579 -0.002631579 -0.002631579
X92attr    0.057894737  0.005263158  0.028947368 -0.023684211  0.028947368
X93attr   -0.005263158 -0.005263158 -0.002631579 -0.002631579  0.050000000
X94attr   -0.005263158 -0.005263158 -0.002631579 -0.002631579 -0.002631579
X98attr   -0.010526316 -0.010526316 -0.005263158 -0.005263158  0.047368421
X100attr   0.042105263 -0.010526316 -0.005263158 -0.005263158 -0.005263158
X103attr  -0.005263158  0.047368421  0.050000000 -0.002631579 -0.002631579
X106attr  -0.005263158 -0.005263158 -0.002631579 -0.002631579 -0.002631579
X126attr   0.036842105 -0.015789474 -0.007894737 -0.007894737 -0.007894737
```

Figure 5. 6: Covariance between users of circle and their attributes

```
> var1=var(my.stddata)
> var1
              X7attr      X14attr      X23attr      X32attr      X43attr
X7attr     0.094736842 -0.010526316 -0.005263158 -0.005263158 -0.005263158
X14attr   -0.010526316  0.094736842  0.047368421 -0.005263158 -0.005263158
X23attr   -0.005263158  0.047368421  0.050000000 -0.002631579 -0.002631579
X32attr   -0.005263158 -0.005263158 -0.002631579  0.050000000 -0.002631579
X43attr   -0.005263158 -0.005263158 -0.002631579 -0.002631579  0.050000000
X50attr    0.021052632 -0.031578947 -0.015789474 -0.015789474  0.036842105
X52attr    0.036842105  0.036842105  0.044736842 -0.007894737 -0.007894737
X53attr    0.047368421 -0.005263158  0.023684211 -0.028947368  0.023684211
X54attr   -0.026315789  0.078947368  0.039473684  0.039473684 -0.013157895
X55attr    0.057894737  0.005263158  0.028947368 -0.023684211  0.028947368
X58attr   -0.005263158 -0.005263158 -0.002631579 -0.002631579 -0.002631579
X59attr    0.036842105 -0.015789474 -0.007894737 -0.007894737 -0.007894737
X60attr   -0.005263158 -0.005263158 -0.002631579 -0.002631579 -0.002631579
X63attr   -0.015789474  0.036842105  0.044736842  0.044736842 -0.007894737
X65attr    0.073684211 -0.031578947 -0.015789474 -0.015789474 -0.015789474
X66attr   -0.005263158 -0.005263158 -0.002631579 -0.002631579 -0.002631579
X68attr   -0.005263158 -0.005263158 -0.002631579 -0.002631579 -0.002631579
X77attr    0.047368421 -0.005263158  0.023684211  0.023684211 -0.028947368
X78attr   -0.047368421  0.005263158 -0.023684211 -0.023684211  0.028947368
X84attr   -0.005263158 -0.005263158 -0.002631579 -0.002631579 -0.002631579
X90attr    0.042105263 -0.010526316 -0.005263158 -0.005263158 -0.005263158
X91attr   -0.005263158 -0.005263158 -0.002631579 -0.002631579 -0.002631579
X92attr    0.057894737  0.005263158  0.028947368 -0.023684211  0.028947368
X93attr   -0.005263158 -0.005263158 -0.002631579 -0.002631579  0.050000000
X94attr   -0.005263158 -0.005263158 -0.002631579 -0.002631579 -0.002631579
X98attr   -0.010526316 -0.010526316 -0.005263158 -0.005263158  0.047368421
X100attr   0.042105263 -0.010526316 -0.005263158 -0.005263158 -0.005263158
X103attr  -0.005263158  0.047368421  0.050000000 -0.002631579 -0.002631579
X106attr  -0.005263158 -0.005263158 -0.002631579 -0.002631579 -0.002631579
X126attr   0.036842105 -0.015789474 -0.007894737 -0.007894737 -0.007894737
X127attr  -0.036842105  0.015789474  0.007894737  0.007894737  0.007894737
X128attr   0.036842105  0.036842105  0.044736842 -0.007894737 -0.007894737
```

Figure 5.7: Variance of social dataset

Step 3: Computation of Eigen values and Eigen vectors from covariance matrix

After finding covariance matrix, the eigenvalues are computed using Equation8.Description of Equation 8 is given in previous chapter. The Eigen value and Eigen vectors of all attribute of profile information of circle0 is given in figure 5.8.

As can be observed from figure 5.8, the Eigenvalues and Eigen vectors are arranged in decreasing order. The first Eigen value is highest Eigen value and last Eigen value is smallest Eigen value of attributes of profile formation. The variance of dataset is equal to total of Eigen value of dataset. The variance and sum of Eigen values is given in figure 5.9. Figure 5.9 shows that total variance of circle0 dataset is equivalent to sum of Eigen values. The attribute names along with their Eigen values and variance is given in Table 5.3. The attributes are arranged according to decreasing order of Eigen values.

```
> # Eigen values and Eigen vectors calculated from covariance matrix
> eigen<-eigen(covariance2)
> eigen
$values
 [1]  1.241971e+00  8.753604e-01  8.200766e-01  5.378652e-01  5.293768e-01
 [6]  4.967103e-01  4.539936e-01  3.666376e-01  3.185543e-01  2.333138e-01
[11]  2.214553e-01  1.875918e-01  1.351862e-01  1.220965e-01  9.333904e-02
[16]  5.041922e-02  1.676987e-02  1.507175e-02  8.038853e-17  6.618602e-17
[21]  5.520305e-17  5.428814e-17  4.468017e-17  3.600871e-17  2.687111e-17
[26]  2.548367e-17  1.936004e-17  1.625108e-17  1.617109e-17  1.334060e-17
[31]  1.310471e-17  1.084108e-17  1.036508e-17  9.816341e-18  9.535742e-18
[36]  9.254534e-18  8.856898e-18  6.618498e-18  5.564902e-18  5.235699e-18
[41]  4.980899e-18  4.870975e-18  4.179371e-18  1.923539e-18  1.881337e-18
[46]  1.169492e-18  1.101511e-18  9.713395e-19  7.197881e-33 -4.090171e-19
[51] -4.764321e-19 -7.773330e-19 -1.172619e-18 -1.560783e-18 -3.552571e-18
[56] -3.755520e-18 -3.870887e-18 -4.454883e-18 -5.417394e-18 -5.873575e-18
[61] -5.925330e-18 -6.345657e-18 -6.399160e-18 -7.163053e-18 -1.032500e-17
[66] -1.204956e-17 -1.528532e-17 -1.705940e-17 -2.448128e-17 -2.882595e-17
[71] -4.390955e-17 -7.843500e-17 -1.479479e-16

$vectors
              [,1]         [,2]          [,3]        [,4]         [,5]          [,6]
 [1,] -0.137858420  0.018017805  1.328275e-01  0.01133074 -0.102759909  0.0139591608
 [2,]  0.067473308  0.007427536  9.935037e-02  0.09493996  0.077187836  0.0898493381
 [3,]  0.005686553  0.010116842  1.213495e-01  0.06708495  0.071919642  0.0755494767
 [4,]  0.044210463  0.043537134  1.758458e-02 -0.01525419  0.007516163 -0.0360197438
 [5,] -0.041623309 -0.054202606 -6.645646e-02  0.02511333 -0.088724765  0.1716040631
 [6,] -0.358768311 -0.014870206 -1.838545e-01  0.07471618  0.009179467  0.0532066446
 [7,]  0.027164485  0.014172486  2.993277e-01  0.24557590  0.079304309  0.1382591353
 [8,] -0.362297012 -0.098696995  1.128556e-01 -0.02400458  0.202695204 -0.0077165524
 [9,]  0.196668336 -0.126492408  1.862733e-01  0.04083839  0.215260636  0.1129835027
[10,] -0.372260222 -0.034077028  7.154461e-02  0.24685733 -0.027841173  0.2020864714
[11,] -0.058157333  0.097471159 -5.945880e-02 -0.05374607  0.186560037  0.1308576551
[12,] -0.206039759 -0.065996395 -9.150093e-02  0.11348419 -0.055337302 -0.1803770546
[13,]  0.016345130 -0.177692477  3.345153e-02 -0.15103152  0.116440143 -0.0061174360
[14,]  0.118536450  0.053889375  1.748210e-01  0.16401489  0.093552298  0.1048010773
[15,] -0.289804041 -0.232010991  1.176036e-01  0.01862422  0.021052514 -0.2579341953
[16,] -0.058157333  0.097471159 -5.945880e-02 -0.05374607  0.186560037  0.1308576551
```

Figure 5.8: Eigen values and Eigen vectors of circle0

```
> #show diagonal values of variance of circle0 cleansed dataset
> diag(variance)
    X7attr    X14attr    X23attr    X32attr    X43attr    X50attr    X52attr    X53attr    X54attr
0.09473684 0.09473684 0.05000000 0.05000000 0.05000000 0.22105263 0.13421053 0.26052632 0.19736842
   X55attr    X58attr    X59attr    X60attr    X63attr    X65attr    X66attr    X68attr    X77attr
0.26052632 0.05000000 0.13421053 0.05000000 0.13421053 0.22105263 0.05000000 0.05000000 0.26052632
   X78attr    X84attr    X90attr    X91attr    X92attr    X93attr    X94attr    X98attr   X100attr
0.26052632 0.05000000 0.09473684 0.05000000 0.26052632 0.05000000 0.05000000 0.09473684 0.09473684
  X103attr   X106attr   X126attr   X127attr   X128attr   X129attr   X133attr   X134attr   X138attr
0.05000000 0.05000000 0.13421053 0.13421053 0.13421053 0.09473684 0.09473684 0.05000000 0.13421053
  X139attr   X141attr   X144attr   X148attr   X149attr   X152attr   X153attr   X156attr   X160attr
0.05000000 0.13421053 0.05000000 0.05000000 0.05000000 0.05000000 0.05000000 0.09473684 0.09473684
  X164attr   X165attr   X169attr   X171attr   X172attr   X173attr   X174attr   X175attr   X179attr
0.05000000 0.09473684 0.09473684 0.09473684 0.05000000 0.09473684 0.05000000 0.05000000 0.05000000
  X181attr   X185attr   X191attr   X192attr   X195attr   X200attr   X201attr   X202attr   X206attr
0.09473684 0.09473684 0.05000000 0.05000000 0.05000000 0.09473684 0.05000000 0.05000000 0.09473684
  X207attr   X210attr   X211attr   X212attr   X213attr   X214attr   X215attr   X216attr   X217attr
0.05000000 0.05000000 0.05000000 0.05000000 0.05000000 0.05000000 0.09473684 0.05000000 0.09473684
  X220attr
0.05000000
> #sum of diagonal values of variance of circle0 dataset
> sum(diag(variance))
[1] 6.715789
> # calculate sum of eigen values of circle0 dataset
> sum(Eigenvalues)
[1] 6.715789
>
```

Figure 5.9: Sum of diagonal variance and eigen values

Table 5.3: Variance of component in Principle Component

Attr_no	Attribute/Dimension	Eigenvalues	Variance %	Cumulative %
Attr92	92 languages;id;anonymized feature 92	1.24E+00	3.87931	3.87931
Attr53	53education;type;anonymized feature 53	8.75E-01	3.87931	7.75862
Attr55	55education;type;anonymized feature 55	8.20E-01	3.87931	11.63793
Attr77	77gender;anonymized feature 77	5.38E-01	3.87931	15.51724
Attr78	78gender;anonymized feature 78	5.29E-01	3.87931	19.39655
Attr65	65 education; year; id;anonymized feature 65	4.97E-01	3.26153	22.65808
Attr50	50 education; school; id;anonymized feature 50	4.54E-01	3.26153	25.91961
Attr54	54education;type;anonymized feature 54	3.67E-01	2.93887	28.85848
Attr52	52 education; school; id;anonymized feature 52	3.19E-01	1.99843	30.85691
Attr59	59 education;year;id;anonymized feature59	2.33E-01	1.99843	32.85534
Attr63	63education;year;id;anonymized feature63	2.21E-01	1.99843	34.85377
Attr126	126locale;anonymized feature 126	1.88E-01	1.99843	36.8522
Attr127	127locale;anonymized feature 127	1.35E-01	1.99843	38.85063
Attr128	128location;id;anonymized feature 128	1.22E-01	1.99843	40.84906
Attr138	138location;id;anonymized feature 137	9.33E-02	1.99843	42.84749
Attr141	141work;employer;id;anonymized feature 140	5.04E-02	1.99843	44.84592
Attr7	7birthday;anonymized feature 7	1.68E-02	1.41065	46.25657
Attr14	14education;concentration;id;anonymized feature 14	1.51E-02	1.41065	47.66722
Attr90	90languages;id;anonymized feature 90	1.27E-16	1.41065	49.07787
Attr98	98languages;id;anonymized feature 98	8.12E-17	1.41065	50.48852
Attr100	100languages;id;anonymized feature 100	5.39E-17	1.41065	51.89917
Attr129	129location;id;anonymized feature 129	5.03E-17	1.41065	53.30982
Attr133	133location;id;anonymized feature 133	4.64E-17	1.41065	54.72047
Attr160	160work;end_date;anonymized feature 157	3.11E-17	1.41065	56.13112
Attr156	156work;employer;id;anonymized feature 52	2.95E-17	1.41065	57.54177
Attr165	165work;end_date;anonymized feature 162	2.92E-17	1.41065	58.95242
Attr169	169work;end_date;anonymized feature 166	2.38E-17	1.41065	60.36307
Attr171	171work;end_date;anonymized feature 168	1.90E-17	1.41065	61.77372

Attr173	173work;end_date;anonymized feature 170	1.89E-17	1.41065	63.18437
Attr181	181work;location;id;anonymized feature 176	1.62E-17	1.41065	64.59502
Attr185	185work;location;id;anonymized feature 177	1.23E-17	1.41065	66.00567
Attr200	200work;position;id;anonymized feature 193	1.23E-17	1.41065	67.41632
Attr206	206work;start_date;anonymized feature 160	9.94E-18	1.41065	68.82697
Attr215	215work;start_date;anonymized feature 201	8.89E-18	1.41065	70.23762
Attr217	217work;start_date;anonymized feature 202	7.39E-18	1.41065	71.64827
Attr23	23education;degree;id;anonymized feature 23	7.24E-18	0.74451	72.39278
Attr32	32education;school;id;anonymized feature 32	6.95E-18	0.74451	73.13729
Attr43	43education;school;id;anonymized feature 43	6.60E-18	0.74451	73.8818
Attr58	58education;year;id;anonymized feature 58	6.19E-18	0.74451	74.62631
Attr60	60education;year;id;anonymized feature 60	5.41E-18	0.74451	75.37082
Attr66	66education;year;id;anonymized feature 66	4.99E-18	0.74451	76.11533
Attr68	68education;year; id;anonymized feature 68	4.56E-18	0.74451	76.85984
Attr84	84hometown;id;anonymized feature 84	4.26E-18	0.74451	77.60435
Attr91	91 languages;id;anonymized feature 91	3.52E-18	0.74451	78.34886
Attr93	93 languages;id;anonymized feature 93	2.82E-18	0.74451	79.09337
Attr94	94languages;id;anonymized feature 94	2.05E-18	0.74451	79.83788
Attr103	103languages;id; anonymized feature 103	1.42E-18	0.74451	80.58239
Attr106	106last_name;anonymized feature 106	9.99E-19	0.74451	81.3269
Attr134	134 location;id;anonymized feature 134	3.23E-19	0.74451	82.07141
Attr139	139 location;id;anonymized feature 138	1.81E-19	0.74451	82.81592
Attr144	144work;employer; id;anonymized feature 143	-2.59E-20	0.74451	83.56043
Attr148	148work;employer; id;anonymized feature 147	-1.02E-19	0.74451	84.30494
Attr149	149work;employer; id;anonymized feature 50	-1.54E-19	0.74451	85.04945
Attr152	152 work;employer; id;anonymized feature 150	-8.92E-19	0.74451	85.79396

Attr153	153work;employer; id;anonymized feature 151	-1.84E-18	0.74451	86.53847
Attr164	164work;end_date; anonymized feature 161	-2.05E-18	0.74451	87.28298
Attr172	172work;end_date; anonymized feature 169	-2.64E-18	0.74451	88.02749
Attr174	174work;end_date; anonymized feature 171	-3.84E-18	0.74451	88.772
Attr175	175work;end_date; anonymized feature 172	-4.63E-18	0.74451	89.51651
Attr179	179work;location; id;anonymized feature 132	-4.73E-18	0.74451	90.26102
Attr191	191 work;position; id;anonymized feature 184	-6.47E-18	0.74451	91.00553
Attr192	192 work;position; id;anonymized feature 185	-7.35E-18	0.74451	91.75004
Attr195	195 work;position; id;anonymized feature 188	-1.09E-17	0.74451	92.49455
Attr201	201work;start_date;anonymized feature 157	-1.25E-17	0.74451	93.23906
Attr202	202work;start_date;anonymized feature 194	-1.49E-17	0.74451	93.98357
Attr207	207 ork;start_date ;anonymized feature 197	-1.73E-17	0.74451	94.72808
Attr210	210work;start_date ;anonymized feature 164	-1.89E-17	0.74451	95.47259
Attr211	211work;start_date;anonymized feature 199	-1.91E-17	0.74451	96.2171
Attr212	212 work;start_date;anonymized feature 165	-2.88E-17	0.74451	96.96161
Attr213	213work;start_date;anonymized feature 166	-3.04E-17	0.74451	97.70612
Attr214	214work;start_date;anonymized feature 200	-3.35E-17	0.74451	98.45063
Attr216	216work;start_date;anonymized feature 168	-3.38E-17	0.74451	99.19514
Attr220	220work;start_date;anonymized feature 171	-1.51E-16	0.74451	99.93965

The attributes ***'92 languages;id;anonymized feature 92','53education;type;anonymized feature 53','55education;type;anonymized feature 55 ', '77 gender; anonymized feature 77', '78gender;anonymized feature 78','65 education; year; id; anonymized feature 65', '50 education; school; id; anonymized feature 50','54education;type;anonymized feature***

54','52 education; school; id; anonymized feature 52'etc have major contribution in making of the circle 0, and the attributes which have zero contribution in making this circle *are* ***'0 birthday; anonymized feature 0', '1 birthday; anonymized feature 1',' 2 birthday; anonymized feature 2',' 48 education; school; id; anonymized feature 48',' 70 education; year; id; anonymized feature 70',' 87 hometown; id; anonymized feature 87',' 86 hometown; id; anonymized feature 86'***, etc.

The study shows that first eighteen attributes explained 46% of total variability of data attributes of profile information. Portion of first four attributes that are Attr92,Attr53,Attr55,Atrr77 are 4%,4%,4%, and 4% as shown in figure 5.10.

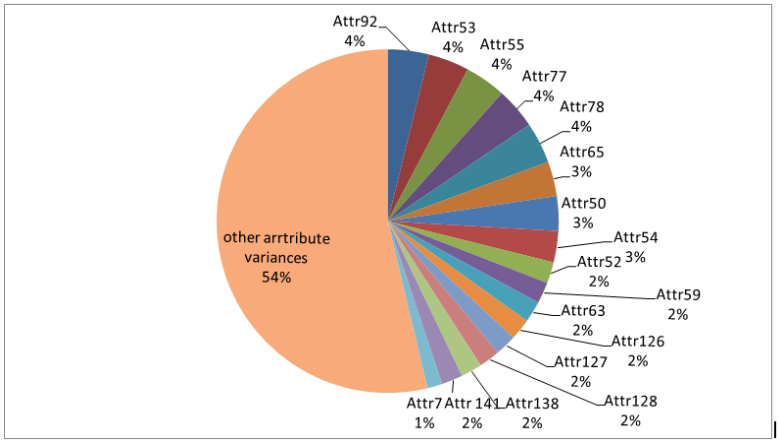

Figure. 5.10: Total variability of data attributes

Step 4: Apply Scree test on Eigen values of attributes of profile information

The scree test plots the eigenvalues or variances with respect to their component, the large eigenvalues/variances and small eigenvalues/variances display "break" between components. The attributes which appear before the "break" are supposed to be meaningful and those attributes which appear after the break are supposed to be unimportant and are not retained. If the scree test shows number of large breaks, in that case attributes appearing before last large break are supposed to be important. The command for plotting scree test in R is given in figure 5.11.

```
> screeplot(my.prc, main="Scree Plot", type="line" )
>
```

Figure 5.11: Command for Scree Plot in R

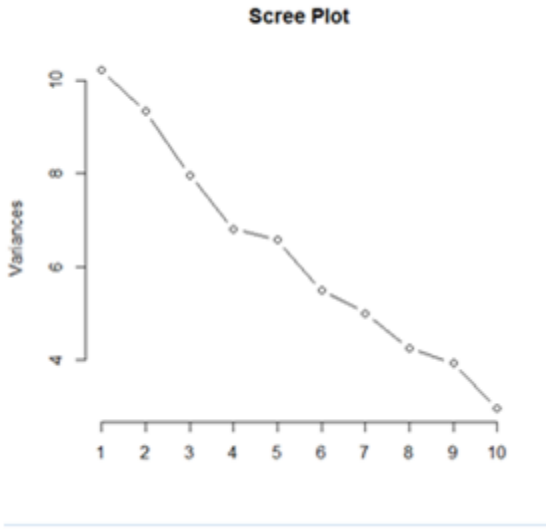

Figure 5.12: Scree Plot of Components according to variance

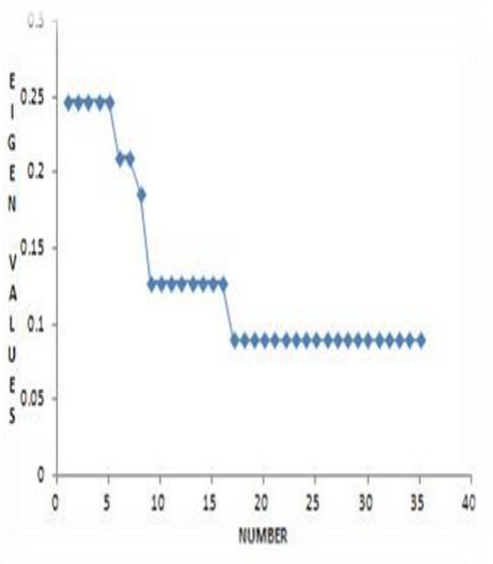

Figure 5.13: Scree Plot of Component with Eigen value

Scree plot of principal component analysis of components with variance in R is shown in figure 5.12, and Scree Plot of components with Eigen values is shown in figure 5.13. Both figures represent number of big breaks before eigenvalues/variance begins to level down. The Scree test shows that first **five components** are significant for the formation of circle.

Step5: Computation of Principal Components

The Eigen vectors of highest Eigen values are Principal Components. The Principal Components are given in figure 5.14.

```
> loadings<-Eigenvectors
> loadings
              [,1]         [,2]          [,3]        [,4]         [,5]
 [1,] -0.137858420  0.018017805  1.328275e-01  0.01133074 -0.102759909
 [2,]  0.067473308  0.007427536  9.935037e-02  0.09493996  0.077187836
 [3,]  0.005686553  0.010116842  1.213495e-01  0.06708495  0.071919642
 [4,]  0.044210463  0.043537134  1.758458e-02 -0.01525419  0.007516163
 [5,] -0.041623309 -0.054202606 -6.645646e-02  0.02511333 -0.088724765
 [6,] -0.358768311 -0.014870206 -1.838545e-01  0.07471618  0.009179467
 [7,]  0.027164485  0.014172486  2.993277e-01  0.24557590  0.079304309
 [8,] -0.362297012 -0.098696995  1.128556e-01 -0.02400458  0.202695204
 [9,]  0.196668336 -0.126492408  1.862733e-01  0.04083839  0.215260636
[10,] -0.372260222 -0.034077028  7.154461e-02  0.24685733 -0.027841173
[11,] -0.058157333  0.097471159 -5.945880e-02 -0.05374607  0.186560037
[12,] -0.206039759 -0.065996395 -9.150093e-02  0.11348419 -0.055337302
[13,]  0.016345130 -0.177692477  3.345153e-02 -0.15103152  0.116440143
[14,]  0.118536450  0.053889375  1.748210e-01  0.16401489  0.093552298
[15,] -0.289804041 -0.232010991  1.176036e-01  0.01862422  0.021052514
[16,] -0.058157333  0.097471159 -5.945880e-02 -0.05374607  0.186560037
[17,]  0.011034363  0.041202977 -9.413846e-05 -0.04262604  0.006254821
[18,] -0.195004943  0.331647870  2.530315e-01 -0.17498682  0.059037488
[19,]  0.195004943 -0.331647870 -2.530315e-01  0.17498682 -0.059037488
[20,]  0.027983039 -0.033143910 -8.041743e-03  0.03874938 -0.102208457
[21,] -0.100109412  0.011677881  1.756530e-01  0.05617156 -0.040050328
[22,]  0.016345130 -0.177692477  3.345153e-02 -0.15103152  0.116440143
[23,] -0.218998522 -0.267684251  2.426810e-01 -0.07655207 -0.254164920
[24,] -0.041623309 -0.054202606 -6.645646e-02  0.02511333 -0.088724765
[25,] -0.052947910  0.007857636  3.356170e-02 -0.01013526 -0.033318503
[26,] -0.025278179 -0.231895083 -3.300493e-02 -0.12591819  0.027715378
[27,] -0.057901147 -0.009838995 -3.528627e-02 -0.09166588 -0.157812159
[28,]  0.005686553  0.010116842  1.213495e-01  0.06708495  0.071919642
[29,] -0.052947910  0.007857636  3.356170e-02 -0.01013526 -0.033318503
[30,] -0.116260272  0.132364489 -1.245146e-01 -0.13748984  0.080225738
[31,]  0.116260272 -0.132364489  1.245146e-01  0.13748984 -0.080225738
[32,]  0.027164485  0.014172486  2.993277e-01  0.24557590  0.079304309
[33,]  0.072193501  0.010393225  9.542837e-03  0.02349519 -0.094692294
```

Figure5.14: Principal Components

According to scree test results first five components are supposed to be meaningful. The structure of first five meaningful components is given in table5.4. The table 5.4 contains only those variables which contribute maximum in principal components.

Each Meaningful Component is explained as below:

Component 1 is eigenvector of first Eigen value. The main parts of first component are 54attr, 78attr, 160attr, 63attr, 127attr. Thus, this component can provide a great grouping among users *(user71, user215, user163, user81, user245, user54, user297, user193, user97, user132, user163, user259)* from the aspect of **54attr,78attr** that is '**54education;type;anonymized feature 54**' and '**78gender;anonymized feature 78'.**

Table 5.4: Structure of first five components for circle0 dataset

Attributes	PC1	PC2	PC3	PC4	PC5
Attr 54	0.196668336	-0.126492408	1.862733e-01	0.04083839	0.215260636
Attr 63	0.118536450	0.053889375	1.748210e-01	0.16401489	0.093552298
Attr 78	0.195004943	-0.331647870	-2.530315e-01	0.17498682	-0.059037488
Attr 127	0.116260272	-0.132364489	1.245146e-01	0.13748984	-0.080225738
Attr 160	0.130426190	-0.002453907	1.388765e-02	0.14003915	0.019384688
Attr 77	-0.195004943	0.331647870	2.530315e-01	-0.17498682	0.059037488
Attr 126	-0.116260272	0.132364489	-1.245146e-01	-0.13748984	0.080225738
Attr 169	-0.105318836	0.101291404	8.263250e-02	0.01256075	0.179828211
Attr 52	0.027164485	0.014172486	2.993277e-01	0.24557590	0.079304309
Attr 92	-0.218998522	-0.267684251	2.426810e-01	-0.07655207	-0.254164920
Attr 128	0.027164485	0.014172486	2.993277e-01	0.24557590	0.079304309
Attr 55	-0.372260222	-0.034077028	7.154461e-02	0.24685733	-0.027841173
Attr 141	0.049460971	-0.029088265	1.699364e-01	0.21724033	-0.094823789
Attr 215	-0.087586775	-0.119795573	-1.707049e-01	0.21454747	0.005695013
Attr 53	-0.362297012	-0.098696995	1.128556e-01	-0.02400458	0.202695204
Attr 165	-0.029618335	-0.243285445	-7.079695e-02	0.03840262	0.210859921

Component 2 is eigenvector of 2nd Eigen value. The second component are affected by attribute **77atrr, 126attr, 169attr**.The second component provide grouping among users (user61, *user229, user81, user253, user264, user29, user110, user183, user334, user245, user* 259,user222) form aspect of **77attr,126attr** that are **'77gender; anonymized feature 77'** and **'126locale; anonymized feature 126'**.

Component 3 is Eigen vector of third Eigen value. The attr54, attr63, attr127, attr77, attr52, attr92, attr128, attr141, attr53 are main part of component 3.Third component provides grouping among users*(user81, user183, user163)* from aspect of attr52, attr128, that are '**52 education; school; id; anonymized feature 52' and'128location; id; anonymized feature 128'.**

Component 4 is Eigen vector of fourth Eigen value. The **attr63, attr78, attr127, attr160,**

attr52, attr128, attr55, attr141, attr215 are main part of component four. Fourth component provides grouping among users *(user81, user163, user183, user297, user229, user193, user29, user132, user222)* from aspect of **' 52 education; school; id; anonymized feature 52',' 128 location; id; anonymized feature 128'** and **'55education; type; anonymized feature 55'**.

Component 5 is Eigen vector of fifth Eigen value. The **attr165, attr53, attr54, attr169** are main part of component five. Fifth component provides grouping among users *(user71, user215, user297, user81, user163, user245)* from aspect of **' 54 education; type; anonymized feature 54'** and **'165work ; end_date; anonymized feature 162'**.

Table 5.5: Representation of major attributes of components with their users,'1' represents user is in group and '0' represents particular user is not in group.

Users	**54attr**	**78attr**	**77attr**	**126attr**	**52attr**	**128attr**	**55attr**	**165attr**
71user	1	1	0	0	0	0	0	1
215 user	1	1	0	0	0	0	0	0
54 user	0	1	0	0	0	0	0	0
61 user	0	0	1	0	0	0	0	0
297 user	0	1	0	0	0	0	1	1
229 user	0	0	1	0	0	0	1	0
81 user	1	0	1	0	1	1	1	0
253 user	0	0	1	0	0	0	0	0
193 user	0	1	0	0	0	0	1	0
97 user	0	1	0	0	0	0	0	0
264 user	0	0	1	0	0	0	0	0
29 user	0	0	1	1	0	0	1	0
132 user	0	1	0	0	0	0	1	0
110 user	0	0	1	0	0	0	0	0
163 user	1	1	0	0	1	1	0	0
259 user	0	1	0	1	0	0	0	0
183 user	0	0	1	0	1	1	1	0
334 user	0	0	1	0	0	0	1	0
245 user	1	0	1	0	0	0	0	0
222 user	0	0	1	1	0	0	1	0

Before applying PCA all 224 attributes are to be considered and which attribute is meaningful in forming the circle is difficult to judge. But after the implementation of PCA only few attributes can be considered in forming the circles.

At the starting of implementation, we have 224 attributes. After the implementation of Principal Component Analysis on data, the numbers of attributes are reduced to only 8 attributes. These 8 attributes are '**54education;type;anonymized feature 54**' ,'**78gender;anonymized feature 78',** '**55 education; type; anonymized feature 55**', '**77 gender; anonymized feature 77**', '**126locale; anonymized feature 126',** '**52 education; school; id; anonymized feature 52,**' '**128location; id; anonymized feature 128', '165work; end_date; anonymized feature 162**' .These attributes are efficient in making a circle.

Firstly, PC1 and PC2 are plotted in figure 5.15, PC1 is x-axis and PC2 is y-axis. Plotting shows that maximum numbers of users are plotted in y-axis direction. Secondly, PC2 and PC1 are plotted in figure 5.16, PC2 is x-axis and PC1 is y-axis. Plotting shows that minimum numbers of users are in y-axis direction. The conclusion is that PC2 is axis in which majority of users vary as compared to PC1.

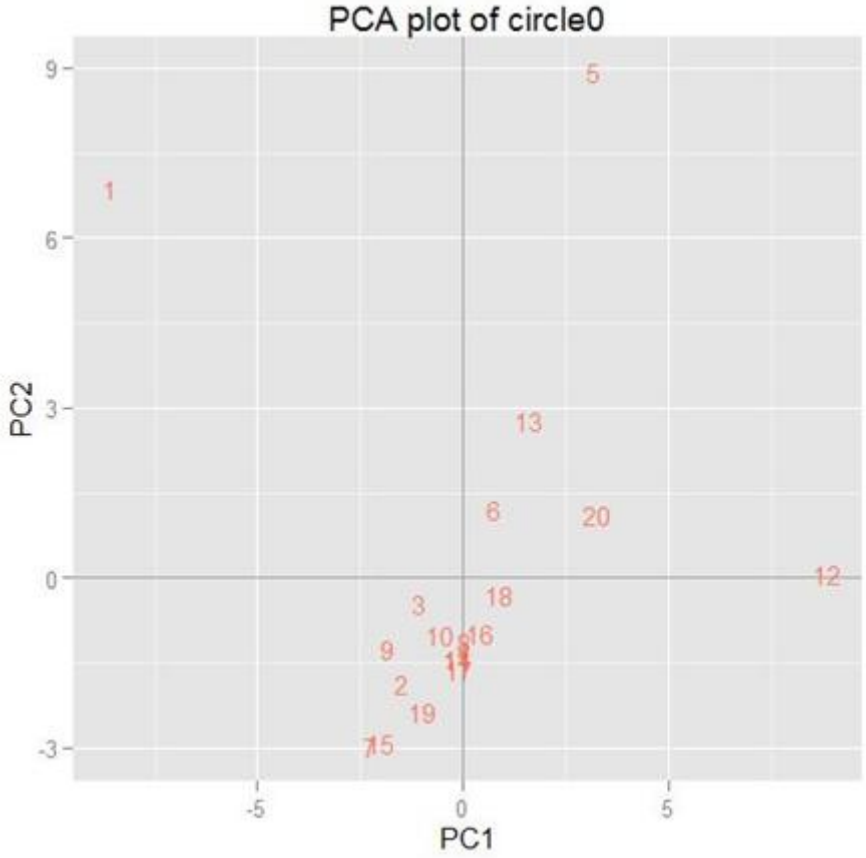

Figure 5.15: PCA plot of circle with PC1 as x-axis

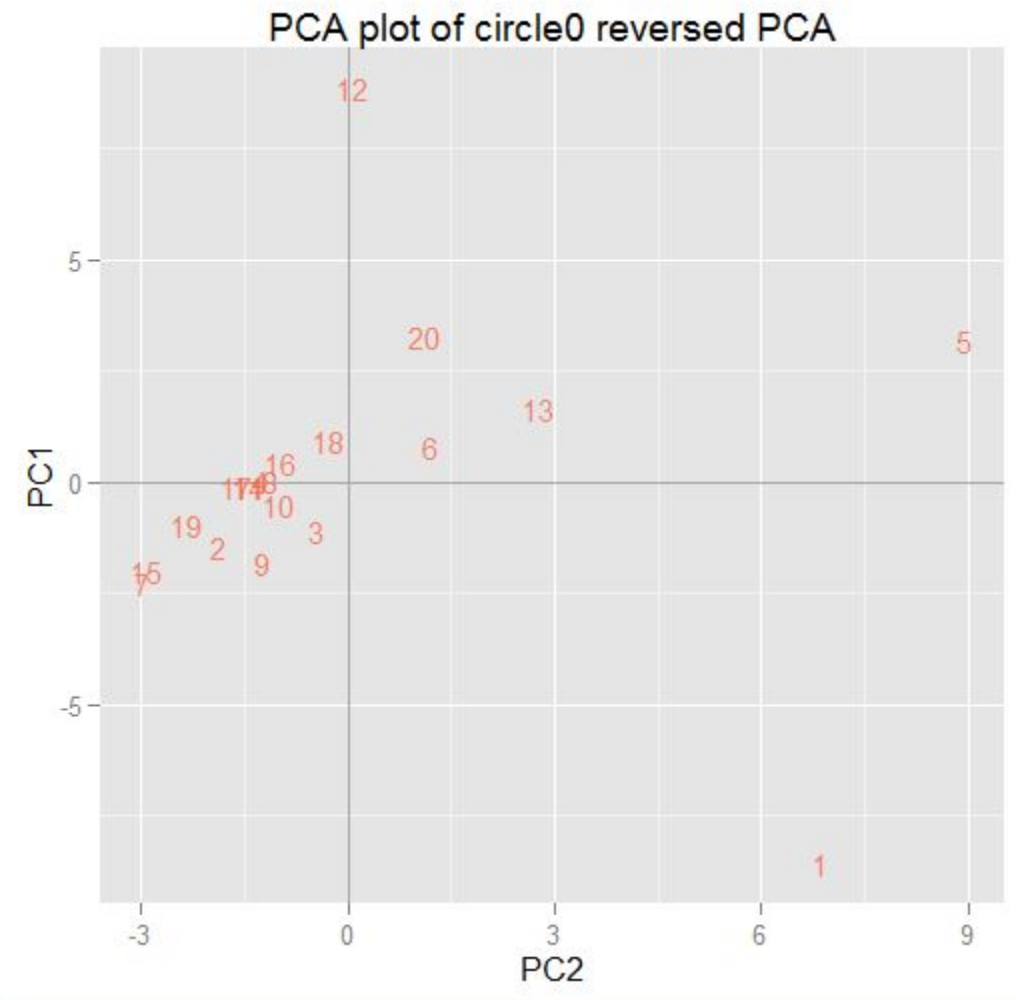

Figure 5.16: PCA plot of circle with PC2 as x-axis

Step 6: Result Interpretation

The PCA after implementation gives eight output attributes which are effective in structuring circle0 of social media dataset. The PC2 is more effective as compared to PC1 is observed from figure 5.15 and figure 5.16. The principal component analysis on attributes of profile information can be observed in figure 5.17.

From figure 5.17, it is observed that fourteen users are between attr77 and attr54 which shows highly correlations among users. One user is between attr54 and atrr165, two users are between attr55 and attr126 and three users are attr165 and attr55. The maximum users are between attr77 and attr54 so **attr77** '**77 gender; anonymized feature 77**' and **attr54**

'**54education; type; anonymized feature 54**' are effective in structuring circle0 in social media data.

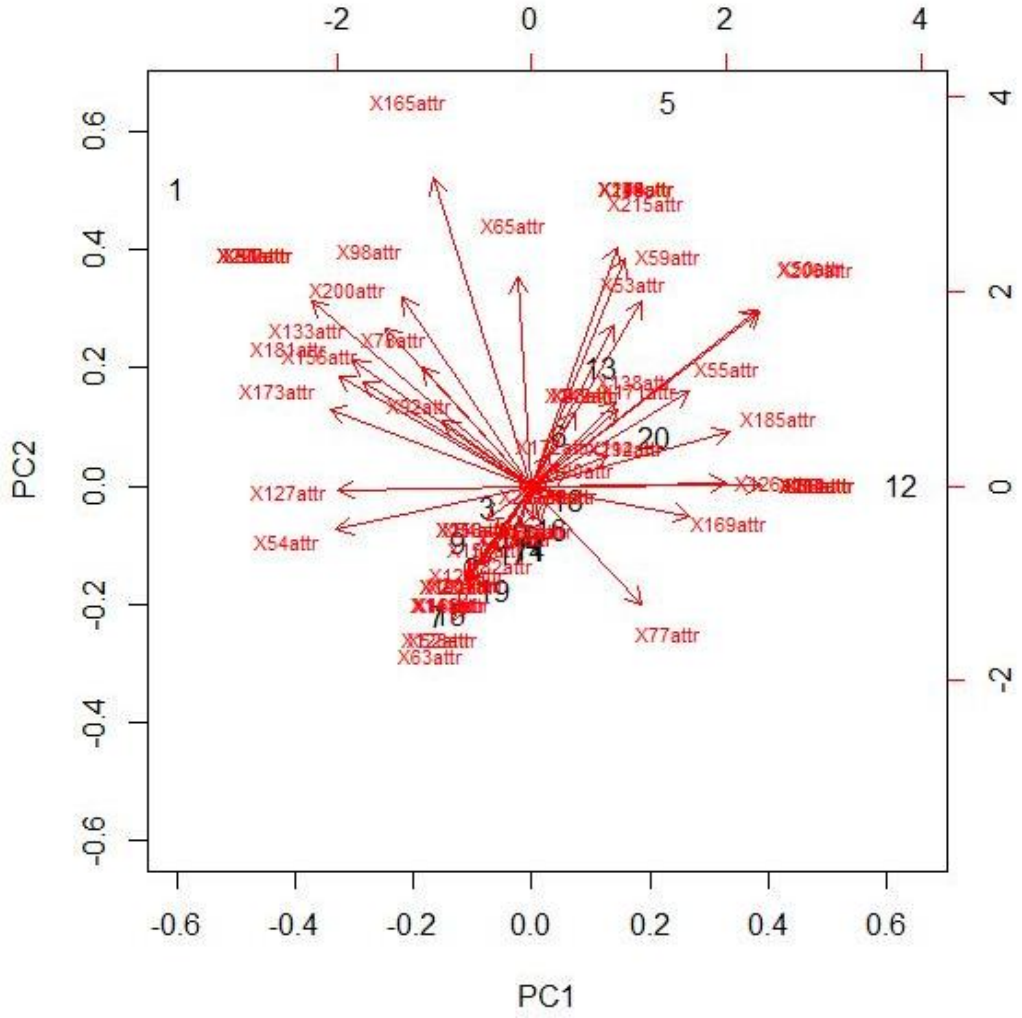

Figure 5.17: PCA Interpretation

CHAPTER 6

CONCLUSION AND FUTURE SCOPE

This chapter is a discussion about the conclusion derived from the research work and its future scope.

6.1 Conclusion:

- Large numbers of data are produced by social networking sites and number of users create and access their accounts daily. The Social networking sites analyze interest of online user and show advertisements on user account according to their interested.
- The social networking site analyses all attributes of users to find similarity and interest. PCA can reduced job of working on all attributes and provides only important attribute which are effective.
- Principal component analysis (PCA) is very powerful mathematical tool for reducing number of dimensions that are very effective in making circle. PCA can reduce effort and cost by reducing unimportant variables.

6.2 Future Scope:

There a few things that can be done to enhance and improvise work done as a part of the social data analysis that is PCA with clustering.

- Future research will focus on analyzing PCA reduced attributes of circles with clustering techniques to understand the effects of PCA in reducing the effort of clustering algorithms.

CHAPTER 7
REFERENCES

Abdi, H., Williams, L. J.,(2010) **" Principal component analysis"**, *Wiley Interdisciplinary Reviews: Computational Statistics, 2*(4), 433-459.

Ahuja, S. P., & Moore, B. (2013) "**State of Big Data Analysis in the Cloud**", *Network and Communication Technologies, 2*(1), p62.

Alt, Rainer, Wittwer and Matthias,(2014) **"towards an ontology-based approach for social media analysis",** Twenty Second European Conference on Information Systems, Tel Aviv .

Bae, Jonghoon, Insead M., (2004) **"Partner substitutability, alliance network structure, and firm profitability in the telecommunications industry**",*Academy of Management Journal* 47.6: 843-859.

Cheng, X., Dale, C., and Liu, J.,(2008) **"Statistics and social network of youtube videos. "**,In Quality of Service, 2008. IWQoS 2008. 16th International Workshop on (pp. 229-238). IEEE (June).

Chin, P. W., Douglas, D. G., Gallagher, E. J., and Yee, B. F. . (2007),U.S. Patent No. 7,191,393. Washington, DC: U.S. Patent and Trademark Office.

Comon , P. ,(1994) **"Independent component analysis, a new concept?"**,*Signal processing, 36*(3), 287-314.

Derakhshan, R., Orlowska, M. E., and Li, X., (2007) **" RFID data management: challenges and opportunities.",** In IEEE International conference on RFID (pp. 175-182) (March).

Ellison, N., (2007) **" Social network sites: Definition, history, and scholarship"**, *Journal of Computer-Mediated Communication*. *13*(1), 210-230.

Goga, O., (2014) **"Matching User Accounts Across Online Social Networks: Methods and Applications"** (Doctoral dissertation, Paris 6).

Hochreiter , R., Waldhauser, C., (2014) **"Data Mining Cultural Aspects of Social Media Marketing. In *Advances in Data Mining. Applications and Theoretical Aspects"***.pp. 130-143. Springer International Publishing.

Ko, M. N., Cheek, G. P., Shehab, M., and Sandhu, R.,(2010) **"Social-networks connect services."** Computer, 43(8), 37-43.

Kolenikov , S., Angeles, G.,(2013) **" The use of discrete data in principal component analysis for socio-economic status evaluation"**, 2005,*Retrieved Nov*, *21*, 2013.

Labrinidis, A., & Jagadish, H. V. (2012) **"Challenges and opportunities with big data**", *Proceedings of the VLDB Endowment*, *5*(12), 2032-2033.

Laurila, J. K., Gatica-Perez, D. (2012)**"The mobile data challenge: Big data for mobile computing research"**, In Pervasive Computing (No. EPFL-CONF-192489).

Leskovec , J. , Krevl, A. ,(2014) **"SNAP Datasets: Stanford large network dataset collection"**.

Lukibisi , F. B., Lanyasunya, T.,(2010) **" Using principal component analysis to analyze mineral composition data"**, In *Proceedings of the 12th kari (kenya agricultural research institue) biennial scientific conference"*, November.pp. 8-12.

Leskovec, J., & Mcauley, J. J. (2012) **" Learning to discover social circles in ego networks.",** In *Advances in neural information processing systems* (pp. 539-547).

Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., & Bhattacharjee, B. (2007, October). "**Measurement and analysis of online social networks**", In*Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*(pp. 29-42). ACM.

Moore, B. ,(1981) **" Principal component analysis in linear systems: Controllability, observability, and model reduction"**, *Automatic Control, IEEE Transactions on 26*(1), 17-32.

Shlens, J.,(2014) **"A tutorial on principal component analysis"**, *arXiv preprint arXiv:1404.1100*.

Smith, G., Boreli , R. , Kaafar, M. A.,(2014) **" A Layered Secret Sharing Scheme for Automated Profile Sharing in OSN Groups"**, In *Mobile and Ubiquitous Systems: Computing, Networking, and Services* .pp. 487-499. Springer International Publishing.

Thakur, B., Mann,M. (2014) **" Data Mining for Big Data"**, International Journal of Advanced Research in Computer Science and Software Engineering Volume 4 Issue 5:469-473.

Wang, Zhan, (2014) **"A big data benchmark suite from internet services."** ,arXiv preprint arXiv:1401.1406.

Wu.X, Zhu.X, Wu.G.Q, Ding.W.,(2014) **"Data mining with big data, *Knowledge and Data Engineering"* IEEE Transactions on *26***(1), 97-107.

CHAPTER 8

LIST OF ABBREVIATIONS

5 V's:-Volume, Velocity, Veracity, Value and Variety.

HDD:-Hard Disk Drive

HDFS:-Hadoop Distributed File System

NoSqL:-Not Only Structured Query Language

RDBMS:-Relational database management system

SSD:-Solid State Drives

PCA:-Principal Component Analysis

PCM:-Phase change Memory

WCC:-Weakly Connected Components

LIST OF PAPERS

Samiya Shafi, Harshpreet Singh, "Identifying Effective Attributes for Structuring Social Circle/Lists in Social Media using Principal Component Analysis ", International Journal of Applied Engineering Research. (Accepted)