A Modified Approach for Data Retrieval for Identifying Primary Causes of Deaths

A H M Noman
Software Engineer, Datos Technologies
The Woodland, Texas
nomancs05@gmail.com

Kumer Das
Assistant Vice President for Research, Innovation, and Economic Development and Assistant Provost, University of Louisiana at Lafayette
kumer.das@louisiana.edu

Stefan Andrei
Professor and Chair, Department of Computer Science
Lamar University, Texas
sandrei@lamar.edu

Abstract

Automatic data retrieval ensures fast, secure and accurate data extraction from structured and unstructured documents. This study introduces a modified approach of retrieving data from the World Health Organization (WHO) database. WHO maintains a large mortality database based on age, sex, and cause of death of various countries all over the world. An Integrated Development Environment (IDE) of programming language R, called RStudio, has been used to develop this retrieval process. There are over 2,000 front-line, user-contributed packages available on the Comprehensive R Archive Network (CRAN), a network of File Transfer Protocol (`FTP`) and web servers around the world. It stores identical, up-to-date, versions of code and documentation for R. The `RCurl` package offers high-level facilities in R to communicate with Hypertext Transfer Protocol (`HTTP`) servers. It is very useful to download Uniform Resource Locators (`URLs`) and submit forms in various ways. R has built-in data frame to store data tables. The `dplyr` package is used to work with the data frames. After data retrieval, the Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) have been applied to find out the major causes of deaths in WHO mortality database. The objective of this research is twofold. Firstly, to retrieve data automatically from large database websites in such a way that it does not rely on any prior knowledge about the target pages and their contents, and secondly, to apply data dimension reduction techniques to identify primary causes of deaths.

Introduction

With the development of computer and database technologies people depend more and more on computers to collect data, process data, and make use of these data. There are some important tools such as machine learning, data mining, knowledge discovery to support us to achieve those tasks (Wyse, Dubes, and Jain, 1980). For example, dimension reduction of data is an effective approach in many different disciplines (Lee, Ciccarello, Acharjee, and Das, 2017). Data retrieval plays an important role to use these tools effectively. Data retrieval means obtaining

the desired data from a Database Management System ("Data Retrieval", 2018). In order to retrieve the coveted data, the user shows an arrangement of criteria by a query. Next, the DataBase Management Systems (DBMS) and programming for managing databases choose the needed data from the databases.

This paper develops a modified approach to retrieve data from large databases. World Health Organization (WHO) mortality database has been chosen for this approach. This database includes some large data files and has over two million data in each of these files. This retrieval approach is capable of extracting required datasets from the WHO mortality database. At that point, the user can select necessary data by executing the queries included in this approach. This process may generate new files with those retrieving data. Finally, the statistical analysis such as Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) on WHO data confirms the feasibility of the approach. The data set in WHO database is disseminated in compressed form. A statistical software R has been used to implement data retrieval process and data analysis. R is an integrated set of software facilities for data manipulation, calculation and graphical display. R is used because it is a well-developed, simple and effective programming language which includes loops, conditionals, user-defined recursive functions and input and output facilities. Moreover, R has a large, intelligible, integrated set of intermediate tools for data analysis.

There exist a numerous number of databases related to the leading causes of deaths in the United States. The Center for Disease Control (CDC) publishes periodically these data in their official website [CDC, 2017], [CDC, 2019].

Our data are extracted from a public database covering a diverse population, but there exist research studies concentrating on a group of people. For example, paper [AML*, 2019] describes the causes of death among homeless people from a perspective of a population-based cross-sectional study of linked hospitalization and mortality data in England.

The primary objective of this research is to obtain data from a huge database website without having a predefined template or structure about the contents. This paper combines the methods currently used in data retrieval, data extraction processes, data analysis techniques, and tools and technologies for managing large datasets. The rest of the paper is organized into four sections: Section 2 introduces a brief description of data retrieval methods; Section 3 describes the data set used in this study and the retrieval techniques used to retrieve the data; Section 4 describes statistical techniques used to model the data; and finally, conclusions appear in Section 5.

Background

Data extraction is an indispensable phase in knowledge discovery process for real-world applications. As a result, it is a very popular feature to practitioners from statistics, pattern recognition, and data mining to machine learning. It is a process that retrieves data from data sources for further processing. Its extraction means to infer a set of new attributes from original attributes through some functional mapping. Suppose there are attributes X_1, X_2, X_3,…, X_n. We have a new set attributes Y_1, Y_2, Y_3,… ,Y_m after extraction where $m < n$, $Y_i = F_i (X_1, X_2, X_3 \ldots X_n)$ and F_i is a mapping function. The objectives of the feature extraction are reducing the amount of data, focusing on the relevant data, and improving the quality of data. There could be two major techniques, such as preprocessing algorithms before data mining, followed by data mining algorithms. In more details, the first approach is to preprocess the data so that it is suitable for data mining (Motoda and Liu, 2002). There exist many algorithms for data extraction. Two of the most popular algorithms for data dimension reduction are the principal component analysis (PCA) and

the singular value decomposition (SVD). PCA is a powerful method for data dimension reduction. It can compress the data by reducing the number of dimensions without much loss of information. In this technique, the original *n* attributes are replaced by another set of *m* new features that are formed from their linear combinations of the original attributes. On the other hand, SVD is a factorization of a matrix via an extension of polar decomposition.

The R language environment has been designed to facilitate the progress of new scientific tools. As R is an open-source programming language, its source code is available for adaptation to other systems. It can be compiled and run on a wide variety of UNIX platforms, Windows and MacOS. The R language can rapidly process large datasets because of continuing improvements in the efficiency of R's coding and memory management. It has extensive and powerful graphics abilities that are closely related with its analytic abilities. The packages in R give access to an up-to-date methodology for leading statistical and other researchers (Maindonald, 2008).
Here there are some useful packages that have been used to retrieve data from WHO database:

- **`RCurl` Package:** The `RCurl` package offers high-level facilities in R to communicate with `HTTP` servers. It is useful to download `URLs` and submit forms in various ways. `RCurl` uses a wrapper for `libcurl` provides function to allow us to compose general `HTTP` requests. It does not only supports `HTTP` and the secure `HTTP`, but also handles authentication using passwords and can use `FTP` to download files. `RCurl` could establish a secure socket connection to write requests to `HTTP` servers and receive the response by implementing the `HTTP` protocol. `RCurl` uses `libcarl`, an existing implementation that is provided in a widely used C library. This is a robust and extensive library that supports `FTP/FTPs/TFTP, SSL/HTTPs, telnet,` as well as `dict`. It also supports cookies, redirects, authentication, etc.
  There are three primary high-level entry points in `RCurl` package. These allow us to fetch a `URL` and submit `HTTP` forms. The functions are `getURL`, `getForm` and `postForm`. A function called `download.file` has been used to download mortality data from WHO database (Lang, 2007).
- **`dplyr` Package:** `dplyr` provides a grammar of data manipulation. It is the next generation of `plyr`, focused on tools for working with data frames. This package has a few set of variables for data manipulation such as `summarise()`, `select()`, `mutate()`, `filter()`, `arrange()`, etc.
- **`sqldf` Package:** It is a package that provides an easy way of performing Structured Query Language (`SQL`) selects on R data frames. The user simply specifies an `SQL` statement in R using data frames in place of table names. It scans the selected statement for tables, performs or accesses a database. `sqldf` supports the `SQLite` backend database, the `PostgreSQL` database, the `H2` java database, and `MySQL`. The `SQLite` and `H2` are embedded serverless zero administration databases. They are included in R driver packages `RSQLite` and `RH2` respectively. As a result, no additional installation is needed for them. On the other hand, `PostgreSQL` and `MySQL` both are client/server database. They must be installed independently. But, they are widely used and very popular (CRAN, 2018 n.d.a.).

The Principal Component Analysis (PCA)

PCA is a multivariate statistical technique that transforms a number of inter-correlated quantitative dependent variables into a set of orthogonal variables called principal components. This procedure can be used to reduce a large set of variables to a small set that still contains most

of the information in the large set. Consider a data matrix with *m* variables and *n* samples where the data are first centered on the means of each variable (Abdi and Williams, 2010).

The first principal component (B_i) can be written by the linear combination of the variables $A_1, A_2, \ldots, A_m$.

Therefore, $B_1 = x_{11}A_1 + x_{12}A_2 + \ldots + x_{1m}A_m$ (Holland, 2008).

Also, consider that B_1 has the highest variance. The weights $x_{11}, x_{12, \ldots,} x_{1m}$ can be calculated with the constraint that their sum of squares is 1.

$$x_{11}^2 + x_{12}^2 + \ldots + x_{1m}^2 = 1$$

Similarly, the second principal component can be calculated where it is correlated with the first component. Let B_2 has the next highest variance. Then,

$$B_2 = x_{21}A_1 + x_{22}A_2 + \ldots + x_{2m}A_m$$

This continues until a total number of *m* principal components equal to the original number of variables. At this time, the sum of the variances of all principal components will equal the sum of the variances of all the variables.

Collectively, $B = XA$.

The calculation of these transformation involves small matrices. The rows of matrix *X* can be considered as the eigenvectors of matrix S_x which are the variance-covariance pairs of the original data. Here, the elements of the diagonal of matrix S_b are known as the eigenvalues. Moreover, the elements of an eigenvector are the weights x_{ij} , known as loadings.

Here, the correlation of variable A_i and principal component Bj for r_{th} sample can be written as

$$r_{ij} = \sqrt{(x_{ij} \vee xr\ (B_j) / s_{ii})}$$

The Singular Value Decomposition (SVD)

SVD is a powerful data reduction technique that can factorize a matrix into a product of three unique components. It uses linear algebra concepts and properties such as matrix and vector computations, normal and orthonormal vectors, determinants, and orthogonality (Boling and Das, 2014).

Consider a rectangular matrix A_{mxn}. The SVD can be defined as follows:

$$A_{mxn} = U_{mxm} S_{mxn} V^T_{nxn}$$

where $U^TU = I$ and $V^TV = I$. This means that *U* and *V* are the orthogonal matrices, *S* is a diagonal matrix in which the nonnegative diagonal elements are singular values of *A*, V^T is the transpose of *V*, the columns of *U* are the orthonormal eigenvectors of AA^T and the rows of *V* are the orthonormal eigenvectors of A^TA. It can be mentioned here that *S* contains the square roots of eigenvalues from *U* or *V* in descending order.

SVD is very useful in the field of text mining and natural language processing. It can identify the data points that show the most variation in the original data items.

Data extraction, retrieval and analysis

Description of Data

This paper documents data extraction process, data retrieval techniques, and statistical analysis of retrieved data. In this regard data has been retrieved from WHO database using R. World Health Organization (WHO) provides a website of WHO mortality database at http://www.who.int/healthinfo/mortality_data/en. This database is a compilation of mortality data

by age, sex and cause of death. It is reported annually by member states from their civil systems. The associated data files of this database can be found from http://www.who.int/healthinfo/statistics/mortality_rawdata/en.

WHO mortality database comprises number of deaths by county, year, sex, age group and cause of death since 1950. The data available in the database are from country civil registration systems. When a death occurs, this system keeps record of it with information of the cause of death. Then national authority compiles the information and submit them to WHO every year (World Health Organization, 2018). WHO verifies that those submitted data are classified according to the International Classification of Diseases (ICD) codes. All non-official codes would be replaced with the most appropriate official codes, if needed.

Explanation of retrieval approach

WHO mortality data files have been stored into the target directory. A user defined function has been used to do that. Here, each subject refers to a file name in the database. `download.file` has been used to download the files from the database. After retrieving the data files, they can be filtered for further statistical analysis. There are different kinds of packages such as `sqldf`, `dplyr` in R to filter the data. The filtered file can be stored as a separate csv file.

```
write.csv(df2, file = "Seychelles_2002_Male.csv")
```

After retrieving data from WHO mortality database into R, each dataset has been split into several modified datasets for analysis. Each modified dataset is an excel file which has been separated from original file based on Country, Year and Sex. For instance, `MortIcd10.zip` file has been chosen for data retrieval process in R.

Statistical Analysis

WHO Data Analysis using PCA

A subset of the WHO database, namely, Japan 2000-Male Data, has been chosen for PCA [Appendix 5]. The total number of deaths considered is 584,970.

Table 1. Calculate Proportion of Variance

Components	Standard Deviation	Proportion of Variance	Cumulative Proportion
PC1	4.920e+04	9.597e-01	9.597e-01
PC2	9.762e+03	3.777e-02	9.975e-01
PC3	2.465e+03	2.410e-03	9.999e-01
PC4	538.30553	0.00011	0.99998
PC5	1.77e+02	1.00e-05	1.00e+00
PC6	78.44	0.00	1.00
PC7	35.15	0.00	1.00

PC8	9.579	0.00	1.00
PC9	6.252	0.00	1.00
PC10	1.765e-12	0.00	1.00

The proportion of variance (Table 5.1) shows that the first component has an importance of 95.9% in predicting the class while the second principal component has an importance of 3.8% and so on which indicates that using the first two components instead of ten features will make our model accuracy to be about 99.8%. Hence, it can be said that 95.8% of data has been explained using only the first principal component where only one-tenth of the entire set of features has been used.

Scree Plot of All Principal Components

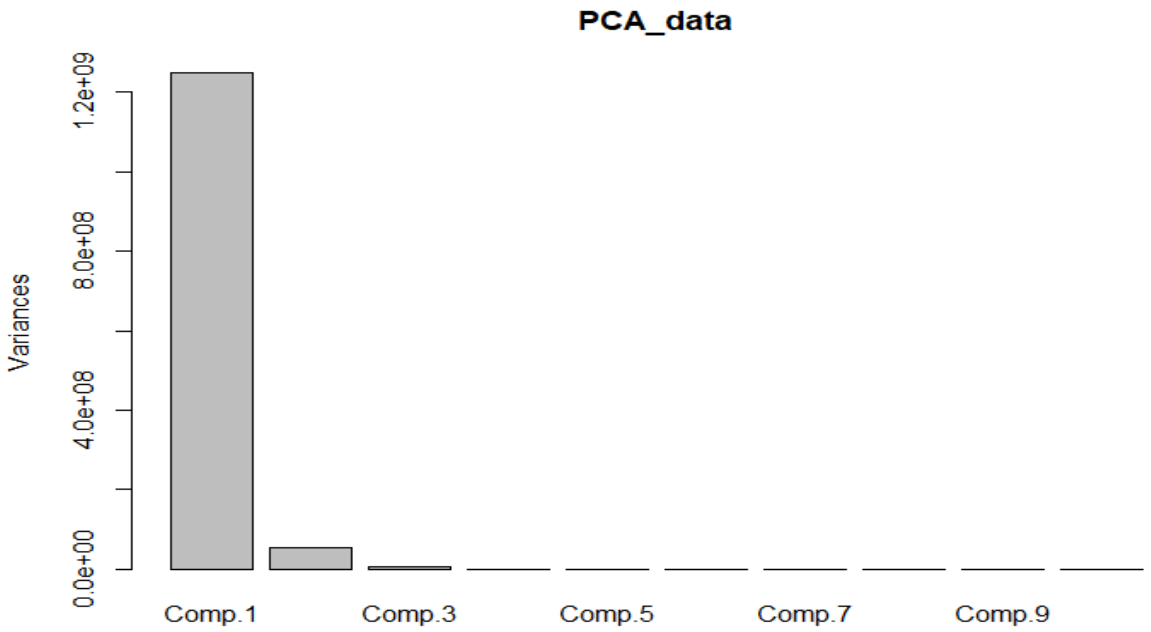

Figure 1. Scree Plot of All Principal Components

The Scree plot in Figure 1 shows all the principal components [Appendix 5]. It is obvious that component 1 is the most important among all. Therefore, it can be said that component 1 and component 2 are good enough to explain entire data set.

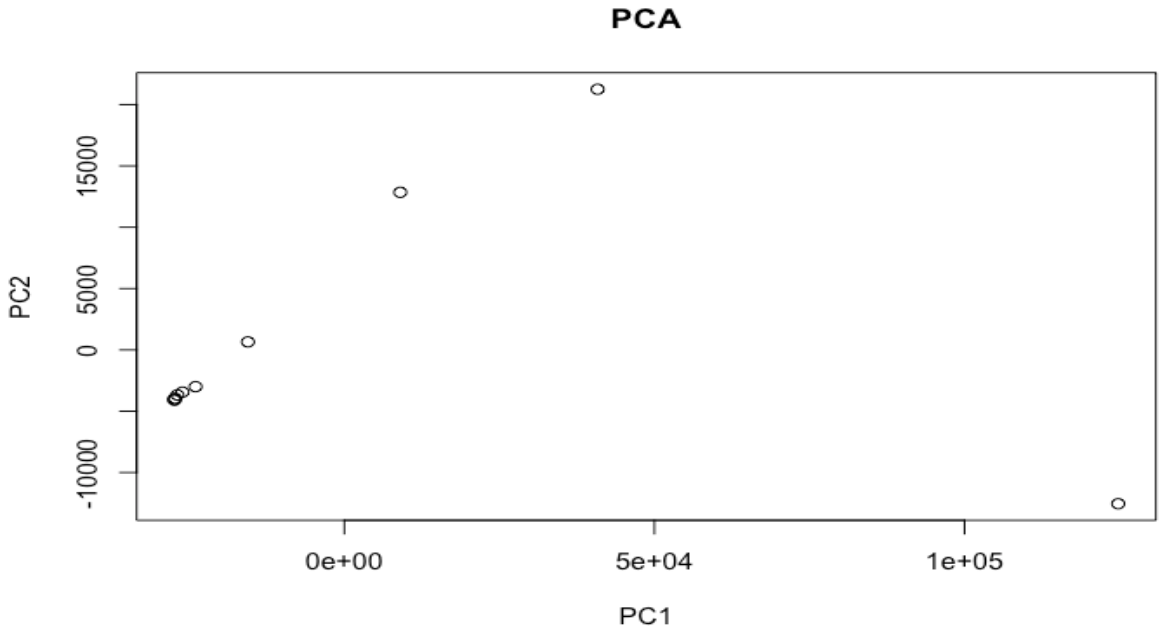

Figure 2. Scatter Diagram of Principal Component 1 (PC1) and PC2

Screen Plot of PCA showing Proportion of Variance

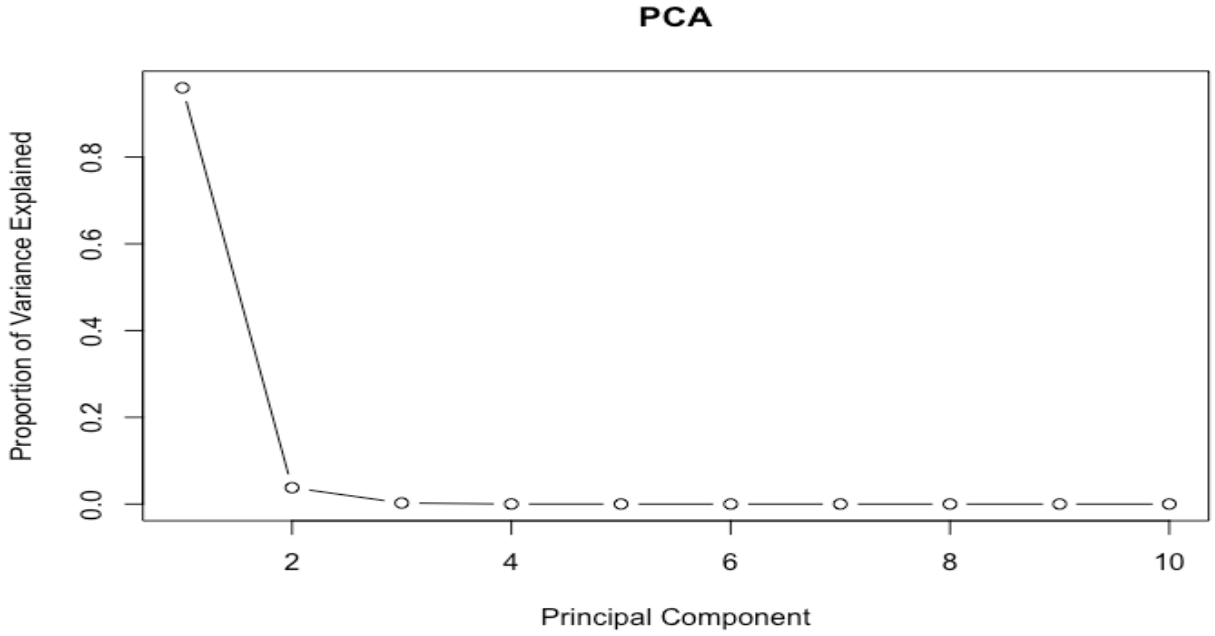

Figure 3. Screen Plot of PCA Showing Proportion of Variance

PCA Biplot

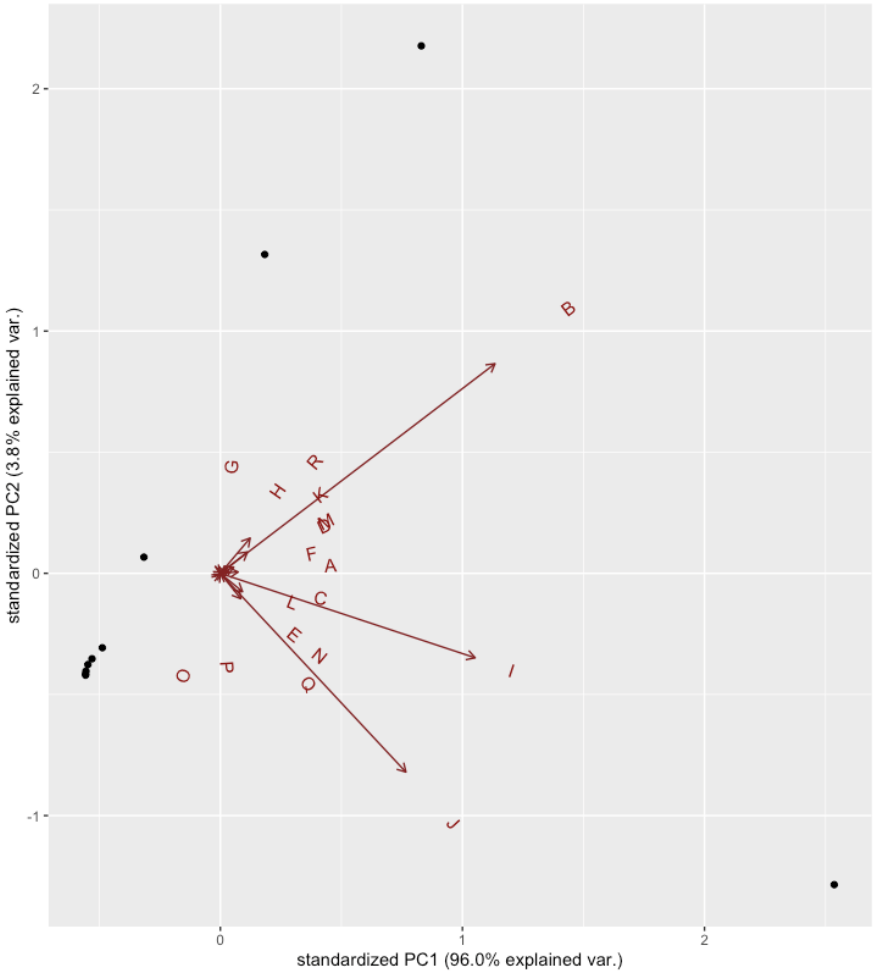

Figure 4. Biplot of PCA

Here, the X-axis and Y-axis represent the first principal component and second principal component, respectively [Appendix 5]. 'Neoplasm', 'Diseases of the Circulatory System', and 'Diseases of the Respiratory System' are almost parallel to the X-axis. That is why, they are combined and almost completely transformed into the first principal component.

Figure 4 shows the arrows as follows: A: Infectious & Parasitic Diseases, B: Neoplasms, C: Diseases of the Blood & Blood Forming Organs, D: Endocrine, Nutritional & Metabolic Diseases, E: Mental & Behavioral Disorders, F: Diseases of the Nervous System, G: Diseases of the Eye & Adnexa, H: Diseases of the Ear & Mastoid Process, I: Diseases of the Circulatory System, J: Diseases of the Respiratory System, K: Diseases of the, Digestive System, L: Diseases of the skin and subcutaneous tissue, M: Diseases of the Musculoskeletal System & Connective Tissue, N: Diseases of the Genitourinary System, O: Certain Conditions Originating in the Perinatal Period, P: Congenital Malformations, Deformations & Chromosomal Abnormalities, Q: Symptoms, Signs & Abnormal Clinical & Laboratory Findings, R: External Causes of Morbidity & Mortality respectively.

It can be seen that the first principal component explains 96% of variance and second principal components explains 3.8%. There are three arrows which are very significant in

composing the first two principal components. They are B: Neoplasm, I: Diseases of the Circulatory System, and J: Diseases of the Respiratory System.

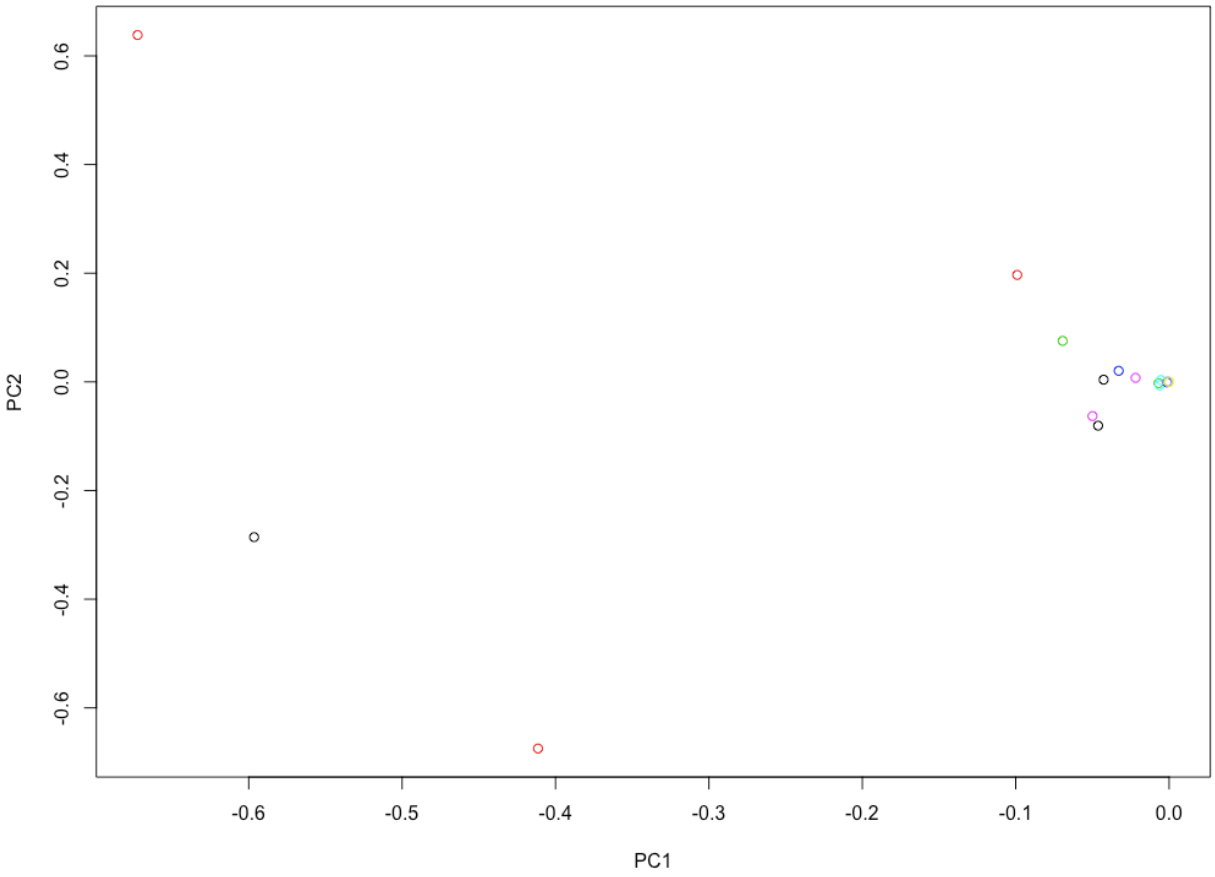

Figure 5. Scatter Diagram of PC1 vs PC2 using SVD

In Figure 5, the X-axis denotes the first principal component and Y-axis denotes the second principal component. Here, most of the points are very close to (0,0) coordinates. There are three points which are very significant for the first principal component and the second principal component since they are scattered in the figure. Text can be added for all the points using the R Code: `text(U[,1],U[,2],colnames (t_mydata),col=cols)` which will justify that the significant three points are B: Neoplasm, I: Diseases of the Circulatory System, and J: Diseases of the Respiratory System. After changing the scale, the same results can be found shown in Figure 6.

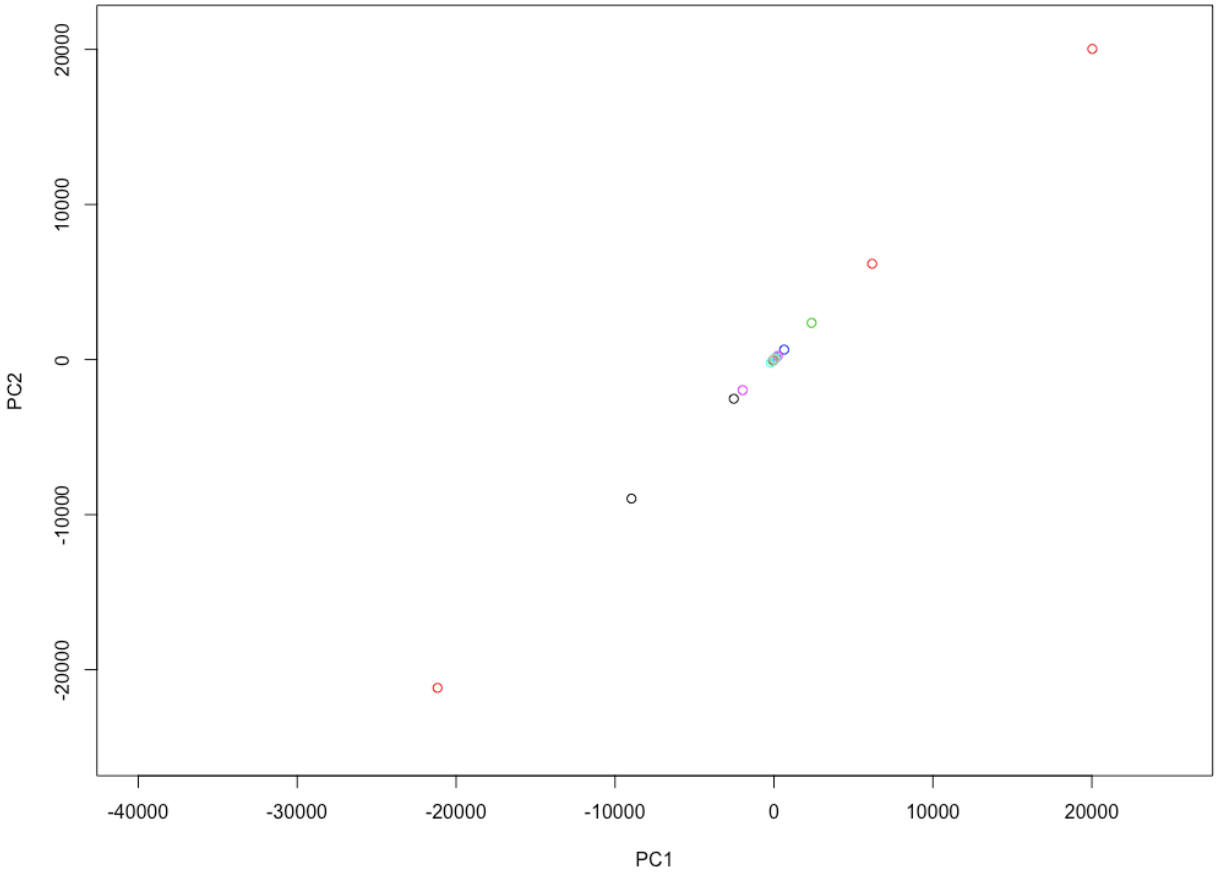

Figure 6. Scatter Diagram of PC1 vs PC2 with Text using SVD

After adding the text using R code: `text (Z[,2], Z[,2], colnames(data), col=cols)` in Figure 6, it can be seen that three points are more scattered than others. These three important points represent different causes of deaths. Since the first principal component and the second principal component are enough to explain approximately 99.8% of causes of deaths, it can be said that 'Neoplasm', 'Diseases of the Circulatory System', and 'Diseases of the Respiratory System' are the main causes of deaths.

The sample R code for this Scatter Diagram as follows:

```
### Scatter diagram (PC1 vs PC2)-Unscaled PC's:
cols<-as.numeric(as.factor(colnames(t_mydata)))
plot(U[,1],U[,2],xlab="PC1",ylab="PC2")
text(U[,1],U[,2],colnames(t_mydata),col=cols)

### Scaled PC's--Graphical presentation using Scattered Diagram:
par(mfrow=c(1,1))
Z<-t(t_mydata)%*%V
plot(Z[,1], Z[,2],xlab="PC1", ylab="PC2")
text(Z[,1], Z[,2], colnames(t_mydata), col=cols)
```

The three important points in Figure 6 represent different causes of deaths such as 'Neoplasm', 'Diseases of the Circulatory System', and 'Diseases of the Respiratory System'. Since the first principal component and the second principal component are enough to explain about 99.8% of causes of deaths, it can be said that 'Neoplasm', 'Diseases of the Circulatory System', and 'Diseases of the Respiratory System' are the main causes of deaths.

Conclusion

This paper shows an approach to retrieve data automatically from WHO mortality database. The R language environment designed to facilitate the progress of new scientific tools has been used to develop this retrieval process. It can be compiled and run on a wide variety of UNIX platforms, Windows and MacOS. Thus, it is possible to retrieve data from a huge database website without having any idea about the contents using this approach. In this study, the `RCurl` package uses a wrapper for `libcurl` that provides functions to allow us to compose general `HTTP` requests. It not only supports `HTTP` and the secure `HTTP`, but also handles authentication using passwords and can use `FTP` to download files. The `sqldf` package provides the way to perform `SQL` selects on R data frames where `dplyr` package has been used to work with additional data manipulation on those data frames. Moreover, R has extensive and powerful graphics abilities. Therefore, applying the powerful methods for statistical analysis such as PCA and SVD prove the feasibility of this study.

References

Abdi, Herve, and Lynne J. Williams (2010). "Principal Component Analysis." *Wiley Interdisciplinary Reviews: Computational Statisics, Vol*. 2(4): 433-459

Boling, Chelsea, and Kumer Das (2015). "Semantic Similarity of Documents using Latent Semantic Analysis." *International Journal of Computer Applications* (09 - 8887), *Vol.* 112 (5): 9-12

CRAN. n.d.a. "Packages: sqldf." Accessed June 03, 2018. https://cran.r-project.org/web/packages/sqldf/README.html

"Data Retrieval." 2018. *Wikipedia*. Last modified August 3, 2017. https://en.wikipedia.org/wiki/Data_retrieval

Holland, Steven M. (2008). "Principal Component Analysis (PCA)." *Department of Geology, University of Georgia, Athens, GA*, 30602-2501.

Lang, Duncan Temple (2007). "R as a Web Client – the `RCurl` package." *Journal of Statistical Software*. *Vol*. VV, Issue II.

Lee, Jaylen, Shannon Ciccarello, Mithun Acharjee, and Kumer Das (2018). "Dimension reduction of gene expression data." *Journal of Statistical Theory and Practice*. 12:2, 450-461. DOI: 10.1080/15598608.2017.1413456

Maindonald, J. H. (2008). "Using R for Data Analysis and Graphics: Introduction, Code and Commentary Centre for Mathematics and Its Applications." *Australian National University*, 96.

Motoda, Hiroshi, and Huan Liu (2002). "Feature selection, extraction, and construction." *Communication of IICM (Institute of Information and Computing machinery, Taiwan), Vol.* 5: 67-72.

World Health Organization. 2018. “Health Statistic and Information System: Download the Raw Data Files of the WHO Mortality Database.” Last modified April 11, 2018. http://www.who.int/healthinfo/statistics/mortality_rawdata/en

Wyse, Neal, Richard Dubes, and Anil K. Jain (1980). “A Critical Evaluation of Intrinsic Dimensionality Algorithms.” *Pattern Recognition in Practice,* 415-425.

Appendix

1. Creating Directory in R

R always pointed at a directory in the computer. It is possible to find out the working directory using `getwd()` function. `getwd()` returns a character string or NULL if the working directory is not available. The `setwd()` function can be used to set up a new working directory. The console output of the above functions is shown as follows:

```
> # Current Working Directory
> getwd()
[1] "C:/Users/ahmnoman/Documents"
> # Set Directory
> setwd("C:/Users/ahmnoman/Desktop/Thesis/TestFiles/Test1")
> myDir <- getwd()
> myDir
[1] "C:/Users/ahmnoman/Desktop/Thesis/TestFiles/Test1"
```

2. Variables in `dplyr` Package

`summarise()`: reduce each group to a smaller number of summary statistics
`select()`: select variables based on their name.
`mutate()`: add new columns
`filter()`: focus on a subset of rows
`arrange()`: change the ordering of the rows

The latest released version can be downloaded from `CRAN`:

```
install.packages("dplyr")
```

The latest development version can be downloaded from github:

```
if (packageVersion ("devtools") < 1.6) {
      install.packages("devtools")
}
devtools: install_github("hadley/lazyeval")
devtools: install_github("hadley/dplyr")
```

3. Download files from WHO Database

```
retrieve_data <- function(targetdir, subject, refresh) {
      filename = paste(subject,".zip",sep="")
      outfile = paste(targetdir,"/", filename,sep="")
      download.file(url=paste("http://www.who.int/entity/healthinfo/sta
                  tistics/",      filename,      "?ua=1",      sep=""),
                  destfile=outfile, method="auto", mode = "wb")
      return(filename)
}
```

4. Filtering Data from Downloaded Files

A partial code for filtering male data in the year 2002 for East African country Seychelles.

```
library(dplyr)
df_filter21 <- filter(df2, Sex %in% c("1"))
df_filter22 <- filter(df_filter21, Country %in% c("1400"))
df_filter23 <- filter(df_filter22, Year %in% c("2002"))
```

The filtered file can be stored as a separate csv file as follows.

```
write.csv(df2, file = "Seychelles_2002_Male.csv")
```

5. PCA in R

As described in Table 1 for Japan 2000-Male data PCA of Japan 2000 data has been applied in R as follows.

```
#Reading data from file
mydata <- read.table("Final_Male_Format_Japan_2000_UPdate.csv",
header=TRUE, row.names=1, sep=",")
#Transpose mydata
t_mydata<-t(mydata)
#Applying PCA
pca_data<-prcomp(t_mydata)
The variables are stored in PCA_data.
```

A partial code for scree plot shown in Figure 3,

```
#scree plot
plot(pca_data$x[, 1], pca_data$x[, 2],main = "PCA", xlab = "PC1", ylab =
"PC2", type = "b")
sDev <- pca_data$sdev
var <- sDev^2
propVar <- var/sum(var)
plot(propVar, xlab = "Principal Component", main = "PCA",
ylab = "Proportion of Variance Explained",
type = "b")
```

Finally, biplot in Figure 4 can be found.

```
biplot(pca_data)
```

6. SVD in R

```
# Data preparation for SVD:
svd_data<-svd(mydata)
U<-svd_data$u
V<-svd_data$v
D<-svd_data$d
qr(mydata)$rank
min(D)
```