Document Clustering Approach Using R Programming with Feedback Data

[1]Dr. Venkata Reddy Medikonda, [2]Dr. Uma Pavan Kumar Kethavarapu

ABSTRACT--*The current day needs of the information technology are based on huge amount of data storage and processing. After storing and processing of such huge amounts of the data such as banking, retail, manufacturing and medical data the fundamental requirement is to analyse that data and getting interesting patterns so as to observe the pulse of the stake holders and identify the challenges to reach the customers or members according to the results of the analysis.The analytics playing a key roleto cater the needs of the various stakeholders andyielding more profits to the companies. The context of analytics can be observed in business point of view and as well as machine learning, data science regards. The current work explains the usage of clustering with R programming packages and methods. Clustering is an unsupervised learning which generates the various clusters for the given data set. A cluster conceptually gives a group of similar data items when compared with other cluster gives opposite properties. In machine learning clustering algorithm plays a vital role in many usecases.The current work considers the source as some document and applying various clustering methods so as to pre-process the given document without unnecessary data. The outcome work is various clusters after processing unnecessary input data from the given input data.*

Keywords: *Clustering, Document clustering, term matrix, analytics, Machine Learning*

I. CLUSTERING USAGE AND IMPORTANCE

Clustering belongs to the class of Unsupervised Learning, where the model was not provided with correct results in the process of training. In Machine Learning system learns from data supplied by user. The following scenarios explain the usage of clustering and the importance of the clustering [1-4].

- A telephone company wants to establish its new towers, for that first the company has to identify the locations where there is usage of bandwidth based on the crowd they have to establish.
- A company wants to construct quarters for the staff by identifying the cluster of the employees in a location and they want to establish a new office there only to improve productivity and reduce the travel time [5].
- A hospital wants to establish chain of hospitals by identifying most accident prone areas so as to serve the lives of the people.

The importance of clustering is to identify the similarities among the elements within the cluster and dissimilarities between the elements of various clusters. A part from these the most important thing is for the huge amounts of the data the possibility of market segmentation, summarization of the news and identifying hidden patterns [7-8].

[1] *Associate Professor, Department of CSE, Narasaraopeta Engineering College, Narasaraopeta, Andhra Pradesh, India*

[2] *Associate Professor, Department of CSE, Malla Reddy Institute of Technology, Hyderabad, Telangana, India*

DOI: 10.37200/IJPR/V24I4/PR201079
Received: 02 Oct 2019 | Revised: 30 Sep 2019 | Accepted: 15 Jan 2020

Sometime the separation of the items are exclusive for a particular to a cluster which is observed in K-Means, in some cases there is possibility of relevancy of element might be belongs to two or more clusters.To identify the dissimilarity among the elements various measures available in the literature of the Machine Learning Algorithms.

- Euclidean distance measure.
- Manhattan Distance Measure.
- Cosine Distance Measure.
- TanimotoDistance Measure.
- Squared Euclidean Distance Measure

II. SCENARIOS OF THE CLUSTERING

The example scenario can be considered like in a class of professional students there is a chance of high attendance and low pass percentage, similarly low attendance and high pass percentage. The same can be observed in case of the students who are not regular and having backlogs might be placed in MNC with highest package so as to identify these kinds of patterns and interesting patterns in the data sets the most suitable method is Clustering.The other dimension of the study is do we get optimal solution all the time, sometimes may stop with local minimum and not with global minimum.The best method of understanding Clustering is with Google page rank algorithm,

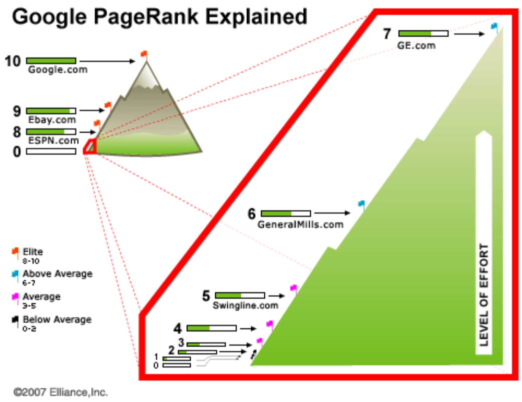

Figure 1: Google Page Rank Observations

The worth of the search engine depends on how fast and accurate results the website gives the above figure shows how Google is expert and efficient in finding the pages according to the requirements of the user given query.

III. DOCUMENT CLUSTERING IMPLEMENTATION

Now will consider the case study of the document clustering, the most common use case in the real time scenarios.The implementation involves, conversion of text into corpus, Convert the text into lower case, removing of the stop words, Removal of punctuations, Removing numbers and extra spaces in the text, creation of the document term matrix and finally generation of the clusters.

DOI: 10.37200/IJPR/V24I4/PR201079
Received: 02 Oct 2019 | Revised: 30 Sep 2019 | Accepted: 15 Jan 2020

```
install.packages("tm")
library(tm)
setwd("C:\\uma")
comments<-read.csv("Comments.csv",head=T)
head(comments)
comments1<-Corpus(VectorSource(comments$Feedback))
comments1
inspect(comments1)
comments2<- tm_map(comments1, tolower)
comments2
comments2[[1]]
comments2[[83]]
comments3<- tm_map(comments2, removeWords, stopwords("english"))
comments3[[83]]
comments4<- tm_map(comments3, removePunctuation)
comments4[[83]]
comments4[[81]]
comments5<- tm_map(comments4, removeNumbers)
comments5[[1]]
comments5[[81]]
comments6<- tm_map(comments5, stripWhitespace)
comments6[[81]]
dtm<- DocumentTermMatrix(comments6)
dtm
inspect(dtm[1:5,1:10])
findFreqTerms(dtm, 10)
findFreqTerms(dtm, 5)
findFreqTerms(dtm, 15)
dtm_tfxidf<- weightTfIdf(dtm)
inspect(dtm_tfxidf[1:5, 1:100])
m <- as.matrix(dtm_tfxidf)
rownames(m) <- 1:nrow(m)
norm_eucl<- function(m) m/apply(m, MARGIN=1, FUN=function(x) sum(x^2)^.5)
m_norm<- norm_eucl(m)
m_norm1<-m_norm[-27,]
cl <- kmeans(m_norm1, 8)
cl
cl$cluster
cl$size
comments_out<-cbind(as.character(comments[-27,]),cl$cluster)
write.csv(comments_out,"Output1.csv")
```

IV. RESULTS

```
findFreqTerms(dtm, 5)
 [1] "media"        "social"        "customers"     "etc"
 [5] "external"     "like"          "sure"          "see"
 [9] "will"         "collaboration" "functionality" "customer"
[13] "great"        "time"          "good"          "integration"
[17] "data"         "outlook"       "can"           "need"
[21] "vendors"
<<DocumentTermMatrix (documents: 5, terms: 100)>>
Non-/sparse entries: 62/438
Sparsity           : 88%
Maximal term length: 16
Weighting          : term frequency - inverse document frequency (normalized) (tf-idf)
Sample             :
    Terms
Docs analytical collaboration    cool  focused functionality mapping
   1    0.00000     0.0000000 0.00000 0.000000      0.000000 0.00000
   2    0.00000     0.0000000 0.00000 0.000000      0.000000 0.00000
   3    0.00000     0.0000000 0.00000 0.000000      0.000000 0.00000
   4    0.00000     0.5580154 0.00000 1.275008      0.000000 0.00000
   5    1.59376     0.0000000 1.59376 0.000000      1.013278 1.59376
```

***Within cluster sum of squares by cluster*:**

```
[1]  6.403549 14.574463  4.594319  5.484033  1.648060  1.746544 25.458124
[8] 10.421510
 (between_SS / total_SS=  12.0 %)
```

Available components:

```
[1] "cluster"      "centers"      "totss"        "withinss"
[5] "tot.withinss" "betweenss"    "size"         "iter"
[9] "ifault"
```

V. CONCLUSIONS

Thisarticle explained the usage of clustering with R programming packages and methods by considering the source as some document and applying various clustering methods so as to pre-process the given document without unnecessary data. The outcome of this work is various clusters after processing unnecessary input data from the given input data.

REFERENCES

1. U. P. K.Kethavarapu, “Various Computing modelsin Hadoop eco system along withthe perspective of analyticsusing R and Machine learning”, International Journal of Computer Science and Information Security, vol. 14, pp. 17-23.
2. U. P. K. Kethavarapu, “The Ten Ingredients of Data Base Systems for Improving Performance and Their Review Leading to Research Problems”, IFRSAs International Journal of computing, vol. 2, no. 2, pp. 409-415, Apr. 2012.
3. S. Madden, “From Databases to Big Data”, IEEE Internet Computing, vol. 16, no.3, 2012.
4. P. Zikopoulos, et. al.,“Understanding bigdata: Analytics forenterprise class hadoop and streamingdata”,McGraw-Hill Osborne Media, 2011.
5. McAfee, et. al., “Bigdata”, Manag. Revolut. Harv. Bus Rev, vol.90,no. 10, pp. 6167, 2012.
6. R. Appuswamy, et. al.,“Scale-up vs Scale-out for Hadoop: Timetorethink?”, In Proc. of 4th Annual Symposium onCloud Computing,2013.
7. C. P. Chen and C.-Y. Zhang, “Data Intensive Applications, Challenges, Techniques and Technologies: A Survey on BigData”, Information Science, vol. 275, pp. 314-347,2014.
8. Lalita Devi, Punam Gaba (2019) Hydrogel: An Updated Primer. Journal of Critical Reviews, 6 (4), 1-10. doi:10.22159/jcr.2019v6i4.33266
9. Fleisig, D., Ginzburg, K., Zakay, D. A model of waiting's duration judgment (2009) NeuroQuantology, 7 (1), pp. 58-65.
10. Song, D. Quantum theory, consciousness, and being (2008) NeuroQuantology, 6 (3), pp. 272-277.

DOI: 10.37200/IJPR/V24I4/PR201079
Received: 02 Oct 2019 | Revised: 30 Sep 2019 | Accepted: 15 Jan 2020