Scientific Journal of Impact Factor (SJIF): 5.71

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 5, Issue 04, April -2018

DIMENSIONALITY REDUCTION FOR BIG DATA AN APPROACH WITH PRINCIPAL COMPONENT ANALYSIS

Ankita Arora [1,] D. Sumathi[2]

[1]M. Tech Student, Department of CSE, Malla Reddy Engineering College, Telangana, India
[2]Professor, Department of CSE, Malla Reddy Engineering College, Telangana, India.

ABSTRACT -*Principal component analysis (PCA) technique is widely used technique for dimensionality reduction in data analysis. There are many benefits to reduce the dimensions of a dataset in different perspective, like visualization of data is restricted to 2 or 3 dimensions. Reducing the dimensions of data can sometimes significantly reduce the time complexity of some numerical algorithms. Besides, most of the statistical models have disadvantage from high correlation between covariates, PCA have the application to produce linear combinations of the covariates that are uncorrelated between each other. Principal component analysis (PCA) uses the concept of orthogonal transformation, which transforms correlated variables of set of observation's to a result set consisting of values which are linearly uncorrelated variables called principal components. The result set we get after this transformation are known as uncorrelated orthogonal basis set.*

KEYWORDS:*Dimensions, Data Analysis, Information retrieval, Sorting Optimization.*

I. INTRODUCTION

Dimensionality reduction is the process of reducing the number of variables which are random in nature, of observations set[1] to get a set of principal variables. This technique can be broadly divided into feature selection and feature extraction.

1.1 FEATURE SELECTION

Feature selection is an approach to find a subset of the original variables (also called features or attributes). There are three methods for this: the filter strategy (e.g. information gain), the wrapper strategy (e.g. search guided by accuracy), and the embedded strategy. It is observed that in some cases, data analysis techniques such as regression or classification works more efficiently when done in the reduced space than in the original space.

1.2 FEATURE EXTRACTION

Feature extraction is another technique to transforms the high-dimensional data space to fewer dimensions space. The data transformation in case of principal component analysis (PCA) is generally linear, but there exist many nonlinear dimensionality reductions too. For high-dimensional data, the tensor representation is useful to reduce the dimensions through multilinear subspace learning.

1.3 PRINCIPAL COMPONENT ANALYSIS

One of the mostly used linear technique to reduce the dimensions, principal component analysis, performs the mapping of the data to a lower-dimensional space in such manner that in the low-dimensional representation the variance of data is maximized. In general, the covariance matrix of the data is constructed and the eigenvectors on this matrix are computed. The eigenvectors that represent the largest eigenvalues (the principal components) can be used to reconstruct a large fraction of the variance of the original dataset. The first few eigenvectors can often be interpreted in terms of the large-scale physical behavior of the system. The original space has been reduced to space spanned by a few eigenvectors, because it is one of the simplest, non-parametric methods to extract relevant information from complicated data sets.
PCA provides a guideline about how to reduce complex dataset to simple dataset with minimum efforts to acknowledge the sometimes hidden and simple result that is difficult to identify in the high dimensional dataset. PCA is one of the most famous multivariate statistical technique and it is useful in almost all types of scientific applications. It helps to analyze a dataset representing observations described by several variables which are dependent on nature, in general, intercorrelated.

II. RELATED WORK

2.1 DIMENSION REDUCTION

For high-dimensional datasets (i.e. with number of dimensions more than 10), dimension reduction is usually performed prior to applying a K-nearest neighbors algorithm (k-NN) in order to avoid the effects of the curse of dimensionality.

The Principal Components can be represented as the following

$$PC\mathrm{i} = \mathrm{a1X1} + \mathrm{a2X2} + \cdots..\mathrm{adXd}$$

where
PCi – Principal Component ‘i’;
Xj – original feature ‘j’;
aj – numerical coefficient for Xj.

Principal Component Analysis in Simple Steps

Principal Component Analysis (PCA) is a very simple yet very famous and useful linear transformation technique widely used in various applications, such as prediction in stock market predictions, gene expression data analysis, and many more. PCA internals can be explained using 3 basic steps.

The headlong size of data is not only a challenge for computer hardware but also a main obstacle for the performance of many machine learning algorithms these days. The main goal of a PCA analysis is to identify patterns in data which are hidden in nature; PCA tries to detect the relationship between variables. If variables are strongly correlated, then only dimensionality reduction is useful. In short, this is what PCA is all about: To find the maximum variance’s direction in high-dimensional data and represent it onto a smaller dimensional subspace without affecting the accuracy.

2.2 PCA Vs. LDA

Both Linear Discriminant Analysis (LDA) and PCA comes under the category of linear transformation methods. PCA yields the directions (principal components) that maximize the variance of the data, whereas LDA also aims to find the directions that maximize the separation (or discrimination) between different classes, which is useful in problem related to pattern classification.

In other words, PCA projects the entire dataset onto a different feature (sub)space, and LDA tries to determine a suitable feature (sub)space in order to distinguish between patterns that belong to different classes.

2.3 PCA and Dimensionality Reduction

Often, the desired aim is to reduce the dimensions of a dd-dimensional dataset by mapping it onto a (k)(k)-dimensional subspace (where k<dk<d) in order to increase the computational efficiency while retaining most of the information. An important question is "what is the size of kk that represents the data 'well'?"

Later, eigenvectors (the principal components) of a dataset are computed and collect them in a projection matrix. Each of those eigenvectors is associated with an eigenvalue which can be interpreted as the "length" or "magnitude" of the corresponding eigenvector. Reduction using PCA onto reduced space is reasonable if there exists eigenvalues having significantly of larger magnitude than others.

III DESIGN OF THE WORKFLOW

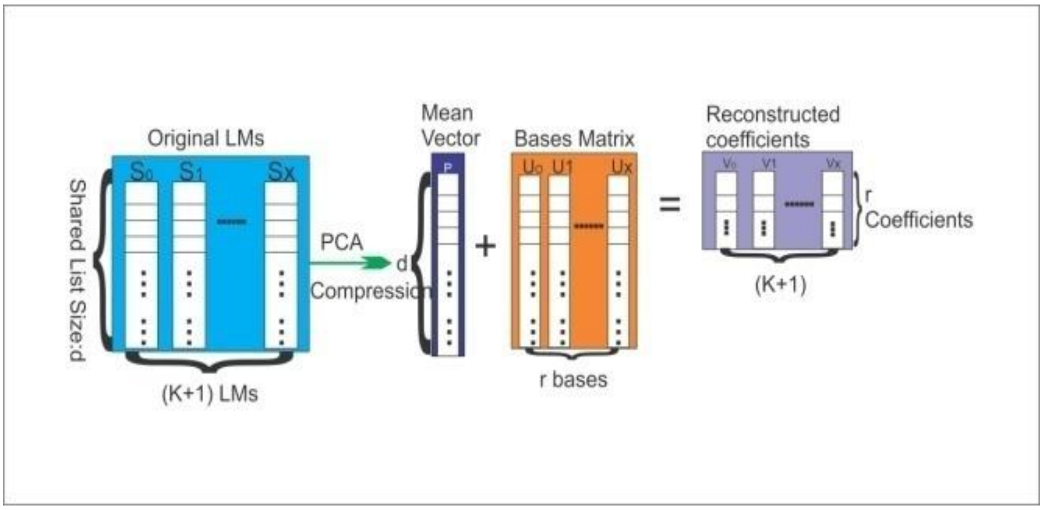


Fig. 3.1(System Architecture)

The system architecture as shown in above fig. explains us such there is a original matrix of the data with the columns starting from S0 to Sx. While applying the statistics and calculate the factors using PCA, a result set has been evolved. This result has been carefully designed with considering some negotiations values which are out of the range from the core matrix, now adding the business required data will generate the efficient results and a reconstructed data set is generated. This data is all power packed with the efficient columns from the original data

IV METHODOLOGY

Principal Component Analysis Dimensional Reduction: (PCADR):
This algorithm approach is to measure data along with principal components rather than on a normal x-y axis. So what are principal components then? They're the underlying structure in the data. They are the directions where there is the most variance, the directions where the data is most spread out.
Steps:

1. Take the dataset.(It can be image or text file.)
2. Convert dataset to matrix P_x where $P_x = [m][n]$ i.e $P_{mxn} = P(x)$
3. Converge the matrix into new matrix with reduced number of rows and columns so that $P^{'} = P(X)$ or $P^{'} = P_{m'xn'}$
4. Identify the linear statistical model.
5. Train the algorithm in such a way that it can find out the results during implementation in a form $P(R) = P_{m''xn:''}$

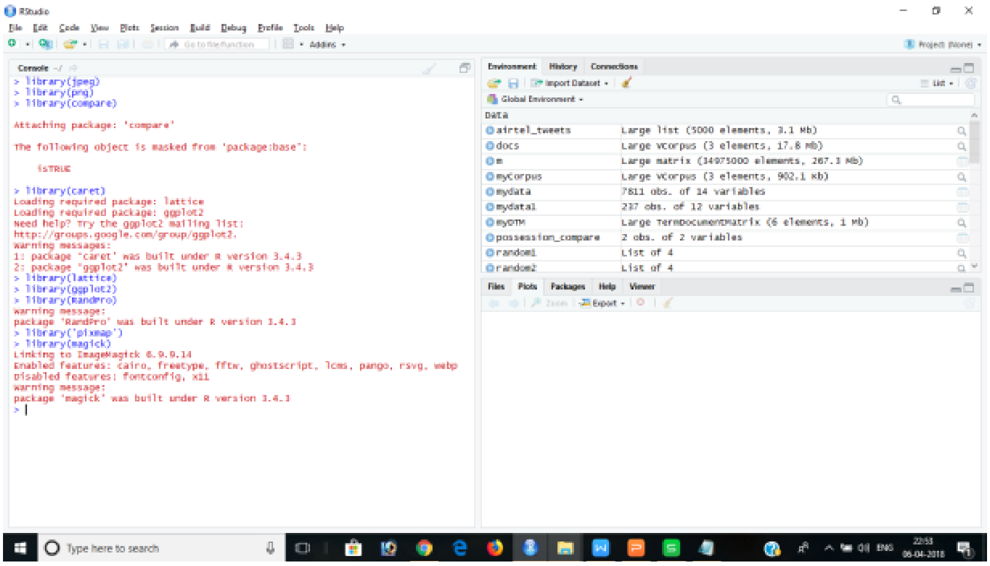
The system has been designed which works with Dimensionality Reduction in analysing the huge amount of data. To do so create a subset of measurements of the data by breaking it down to different levels of sets. This is achieved by analyzing the principal components of the input variables. The analysis of PC is valuable when attributes are highly correlated in different set of measurement. In that case it provides a few (often less than three) variables that are weighted combinations of the original variables that retain the explanatory power of the full original set
library(jpeg) -- used for jpeg image processing library(png) - used for png image processinglibrary(compare) - for estimating the comparative between two methodologieslibrary(caret) - lodes algorithm from binary search to SVM - includes PCAlibrary(ggplot2) - for plotting the resultslibrary(RandPro) - Random Projectionlibrary('pixmap') - image manipulationslibrary(magick) - image manipulations

Chunks of Data
After accessing the image do chunking of data. While storing images, it need to be stored in different formats for better analysis.

```
chunk_Image<-function(image,granularity=10,result_length=100){
img_id<-(image_1[,,1]+image_2[,,2]+image_3[,,1])/3
  }
```

Data Reduction
To reduce the image dimension we use PCA models, this helps us in reducing the dimension values of the images without effecting the fore ground of the image. The concept is the background of the image pixel values are eliminated from the image - considering the brighter pixels with more intensity are being eliminated from the image.

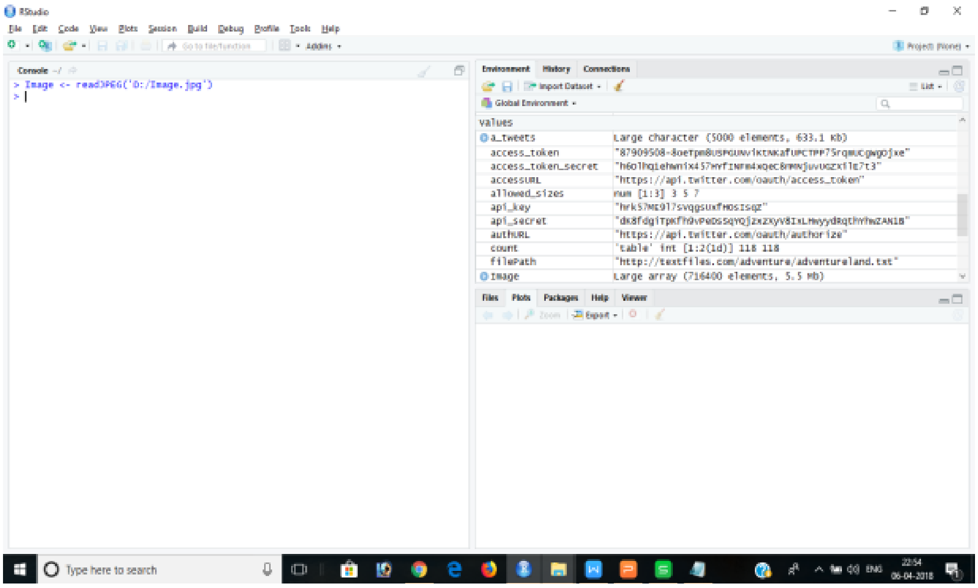
V. EXPERIMENT RESULTS



```
> library(jpeg)
> library(png)
> library(compare)

Attaching package: 'compare'

The following object is masked from 'package:base':

    isTRUE

> library(caret)
Loading required package: lattice
Loading required package: ggplot2
Need help? Try the ggplot2 mailing list:
http://groups.google.com/group/ggplot2.
Warning messages:
1: package 'caret' was built under R version 3.4.3
2: package 'ggplot2' was built under R version 3.4.3
> library(lattice)
> library(ggplot2)
> library(RandPro)
Warning message:
package 'RandPro' was built under R version 3.4.3
> library('pixmap')
> library(magick)
Linking to ImageMagick 6.9.9.14
Enabled features: cairo, freetype, fftw, ghostscript, lcms, pango, rsvg, webp
Disabled features: fontconfig, x11
Warning message:
package 'magick' was built under R version 3.4.3
> |
```

```
Data
airtel_tweets        Large list (5000 elements, 3.1 Mb)
docs                 Large VCorpus (3 elements, 17.8 Mb)
m                    Large matrix (34975000 elements, 267.3 Mb)
myCorpus             Large VCorpus (3 elements, 902.1 Kb)
mydata               7611 obs. of 14 variables
mydata1              237 obs. of 12 variables
myDTM                Large TermDocumentMatrix (6 elements, 1 Mb)
possession_compare   2 obs. of 2 variables
random1              List of 4
random2              List of 4
```

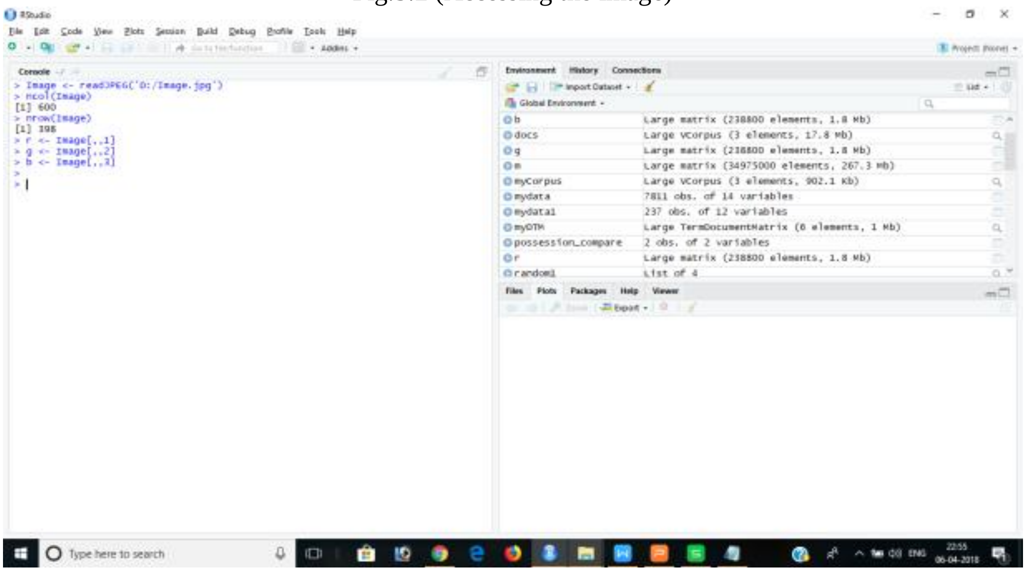
Fig.5.1 (Libraries to console)



```
> Image <- readJPEG('D:/Image.jpg')
> |
```

```
values
a_tweets              Large character (5000 elements, 633.1 Kb)
access_token          "87909508-8oeTpm8USPGUNv1KtNKafUPCTPP75rqmUCgwgOjxe"
access_token_secret   "h6o1hqiehwnix457HYfINFm4xqec8mMNjuvUGzxi1E7t3"
accessURL             "https://api.twitter.com/oauth/access_token"
allowed_sizes         num [1:3] 3 5 7
api_key               "hrk57ME917svqgsuxfHOSIsqz"
api_secret            "dk8fdgiTpKfh9vPeDSSqYQjzxZxyV8IxLHwyydRqthYfwZAN1B"
authURL               "https://api.twitter.com/oauth/authorize"
count                 'table' int [1:2(1d)] 118 118
filePath              "http://textfiles.com/adventure/adventureland.txt"
Image                 Large array (716400 elements, 5.5 Mb)
```

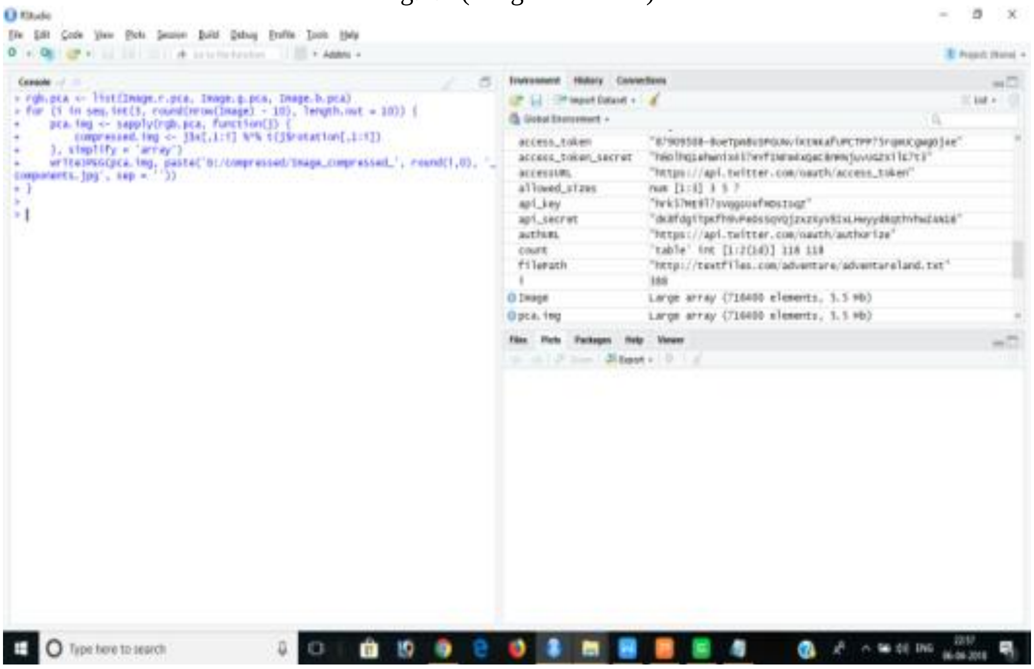
Fig.5.2 (Accessing the image)



```
> Image <- readJPEG('D:/Image.jpg')
> ncol(Image)
[1] 600
> nrow(Image)
[1] 398
> r <- Image[,,1]
> g <- Image[,,2]
> b <- Image[,,3]
>
> |
```

```
b                     Large matrix (238800 elements, 1.8 Mb)
docs                  Large VCorpus (3 elements, 17.8 Mb)
g                     Large matrix (238800 elements, 1.8 Mb)
m                     Large matrix (34975000 elements, 267.3 Mb)
myCorpus              Large VCorpus (3 elements, 902.1 Kb)
mydata                7811 obs. of 14 variables
mydata1               237 obs. of 12 variables
myDTM                 Large TermDocumentMatrix (6 elements, 1 Mb)
possession_compare    2 obs. of 2 variables
r                     Large matrix (238800 elements, 1.8 Mb)
random1               List of 4
```

Fig.5.3 (Image to matrix)



```
> rgb.pca <- list(Image.r.pca, Image.g.pca, Image.b.pca)
> for (i in seq.int(3, round(nrow(Image) - 10), length.out = 10)) {
+     pca.img <- sapply(rgb.pca, function(j) {
+         compressed.img <- j$x[,1:i] %*% t(j$rotation[,1:i])
+     }, simplify = 'array')
+     writeJPEG(pca.img, paste('D:/compressed/Image_compressed_', round(i,0), '_
components.jpg', sep = ''))
+ }
>
> |
```

```
access_token          "87909508-8oeTpm8USPGUNv1KtNKafUPCTPP75rqmUCgwgOjxe"
access_token_secret   "h6o1hqiehwnix457HYfINFm4xqec8mMNjuvUGzxi1E7t3"
accessURL             "https://api.twitter.com/oauth/access_token"
allowed_sizes         num [1:3] 3 5 7
api_key               "hrk57ME917svqgsuxfHOSIsqz"
api_secret            "dk8fdgiTpKfh9vPeDSSqYQjzxZxyV8IxLHwyydRqthYfwZAN1B"
authURL               "https://api.twitter.com/oauth/authorize"
count                 'table' int [1:2(1d)] 118 118
filePath              "http://textfiles.com/adventure/adventureland.txt"
i                     388
Image                 Large array (716400 elements, 5.5 Mb)
pca.img               Large array (716400 elements, 5.5 Mb)
```

Fig.5.4 (Dimesion Reduction)

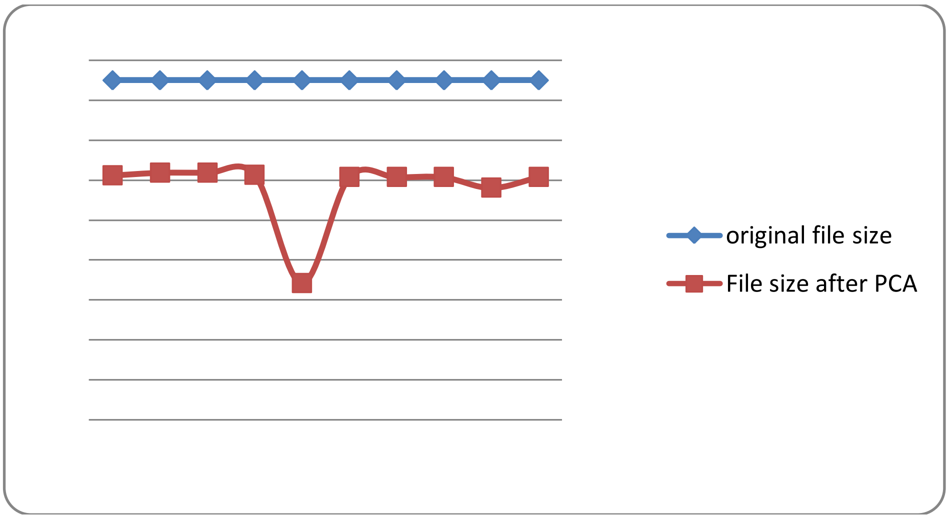


Fig. 5.5(File sizes after applying PCA)

VI. CONCLUSION

From the above results we can specify that usage of PCA is comparatively better than the Random Projection. The PCA works well on the images and this can be used to reduce the size of the image comparatively by reducing the dimension in the point of reduction the brightness of the pixies and converting the image size to very less from its original size.

VII. FUTURE ENCHANCEMENT

In future, we will continue working on it and refine results with low computation time also. We propose new mechanism in which all modules are combined like Reduction with different methodologies which can even helps e commerce organization works with machine learning in reducing the size of the image. This helps in increasing their business needs and leads to increase in the revenue.

VIII REFERENCES

[1] Lu Z X. Research and Improvement of PageRank Sort Algorithm Based on Retrieval Results International Conference on Intelligent Computation Technology and Automation. IEEE, 2014:468 - 471.

[2] Michal Cutler, Yungming Shih, Weiyi Meng. “Using the structures of HTML documents to improve retrieval”, in Proc of Usenix Symp on Internet Technologies and Systems (USITS’ 97). Piscataway, NJ: IEEE, 1997: 241- 251.

[3] Cutler M, Deng H, Maniceam S S, et a1. “A new study on using HTML structures to improve retrieval”,in Proc of the 11th IEEE Conf on Tools with Artificial Intelligence (ICTAl). Piscataway, NJ: IEEE, 1999: 406-409.

[4] Lawrence Page, Sergey Brin, Rajeev Motwani. “The PageRank citation ranking: Bring order to the Web”, S1DL-WP-1999-0120. Stanford: Stanford lnfoLab Publication Server, 1999.

[5] Sergey Brin, Larry Page. “The Anatomy of a large-scale hypertextual Web search engine”, Computer Networks and ISDN Systems(ISDN), 1998, 30: 107-1 17.

[6] Jon M Kleinberg. “Authoritative sources in a hyperlinked environment”, in Proc of ACM- SIAM Symposium on Discrete Algorithms. New York: ACM, 1998: 668-677

[7] Han X, Li J, Gao H. Efficient Top-k Retrieval on Massive Data. IEEE Transactions on Knowledge & Data Engineering, 2015, 27(10):1-1.

[8] Bharat K, Mihaila G A. When Experts Agree: Using Non-affiliated Experts to Rank Popular Topics[C]. Proceedings of the 10th International Conference on World Wide Web. ACM, 2001: 597-602.

[9] http://baike.baidu.com/view/1518.html?fromTaglist.

[10] Shaohua L, Wenyu G. Survey of Page-ranking Algorithms[J]. Application Research of Computers, 2007, 24(6): 4-7.

[11] McSherry F. A Uniform Approach to Accelerated PageRank Computation[C]. Proceedings of the 14th International Conference on World Wide Web. ACM, 2005: 575-582.

[12] Eugene Agichtein, Eric Brill, Susan Dumais, et al. “Learning user interaction models for predicting Web search result”, in Proc of the ACM Conf on SIGIR. New York: ACM, 2006: 3-10

[13] Financial Hot Words http://app.xinhua08.com/tags.php

[14] Stojanovic, Branka, and Aleksandar Neskovic. "Impact of PCA based fingerprint compression on matching performance." In Telecommunications Forum (TELFOR), 2012 20th, pp. 693-696. IEEE, 2012.

[15] Beltran, Luis A. "Nonparametric multivariate statistical process control using principal component analysis and simplicial depth." PhD diss., University of Central Florida Orlando, Florida, 2006.