IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 3, March 2016

Development of an Excellent System for Information Retrieval Corresponding to Vivid Comparisons of Distinct Methodologies Incorporating DR and Clustering Schemes

Mr. Aditya G. Bhor[1], Ms. Komal K. Rahane[2], Ms. Poonam P.Rajurkar[3], Ms. Neha S. Pathak[4], Ms. Anagha N. Chaudhari[5]

Department of Information Technology, Pimpri Chinchwad College of Engineering, Pune, India[1,2,3,4]

Asst. Professor, Department of Information Technology, Pimpri Chinchwad College of Engineering, Pune, India[5]

Abstract: We are accompanied by huge data nowadays. Everyone produces enormous data through variety of operations, transactions and devices. Ultimately it ends up with the overhead for machines to generate and keep such huge data. The noticeable exponential growth of data becomes difficult and utmost challenging. Such data is literally large and not easy to work with for storage and retrieval[15]. This type of data can be treated with various efficient techniques for cleaning, compression and sorting of data[15]. Pre processing can be used to remove basic English stop-words from data making it compact and easy for further processing; later dimensionality reduction techniques make data more efficient and specific[16]. This data later can be clustered for better information retrieval[16]. This paper elaborates the various dimensionality reduction and clustering techniques applied on sample dataset C50test of 2500 documents giving promising results, their comparison and better approach for relevant information retrieval.

Keywords: High Dimensional Datasets, Dimensionality reduction, SVD, PCA, Clustering, K-means, Fuzzy Clustering, Hierarchical Clustering.

I. INTRODUCTION

Working with huge and high dimensional data is a tedious task. Management of data is very necessary; and the basic step is to prepare data for processing. This stage is known as preprocessing, done with stop word removal and stemming. The data obtained for processing contains enormous attributes and thus complexity is increased, this can be resolved by using dimensionality reduction approaches like SVD and PCA. When the data is ideal for information retrieval it can be just organized into clusters and then efficient and promising results of informational retrieval can be obtained. Clustering has various approaches as we implemented like K-means, Fuzzy Clustering and Hierarchical clustering.

Modules

Module 1: Data Preparation

Module 2: Comparison of results of various Dimensionality Reduction (DR) techniques

Module 3: Comparison of effective approach for clustering and development of Information Retrieval (IR) System.

Module 1: Preprocessing

Data pre-processing is an important step in the data miningprocess[17]. The phrase "garbage in, garbage out" is particularly applicable to data mining and machine learningprojects[2]. If there is much irrelevant and duplicate information present or noisy and unreliable data, then knowledge discoverygets more difficult[2].

II. SYSTEM ARCHITECTURE

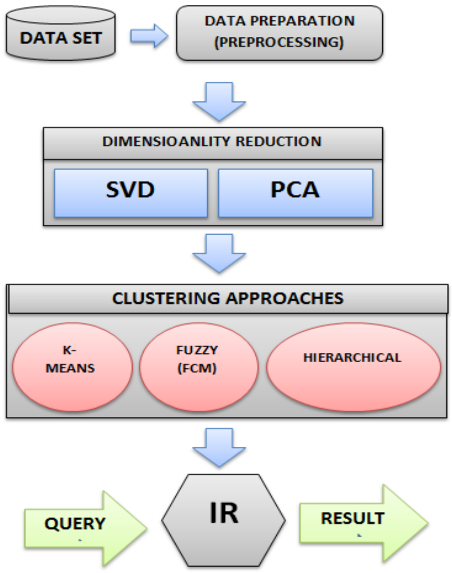


Fig.1.System Architecture

As there are many stop words in a given text file at any given instance, these words increase the dataset size and also slows the further processing of data mining techniques [1].Herein use of stop word and stemming is done for preparation of data i.e. the preprocessing. The purpose of both this method is to remove various suffixes, to reduce number of words, to have exactly matching stems, to save memory space and time[5].

DOI 10.17148/IJARCCE.2016.5381

Module 2 :Application of Dimensionality Reduction Techniques

DR techniques are proposed as a data preprocessing step. This process identifies a suitable low dimensional representation of previous data[3]. Dimensionality Reduction (DR) in the dataset improves thecomputational efficiency and accuracy in the analysis of data[15].Dimensionality reduction is the process of reducing the number of random variables under some consideration. A word matrix (documents*terms) is given as input to reduction techniques like Principal Component Analysis (PCA) and Singular Value Decomposition (SVD)[17].

A. Singular Value Decomposition(SVD)

SVD can be implemented using formula:

$$\mathbf{A}_{[m \times n]} = \mathbf{U}_{[m \times k]} * \sum_{[k \times k]} * (\mathbf{V}_{[k \times n]})^{T}$$

where,
A:*m x n* matrix (m documents, n terms)
U:*m x k* matrix (m documents, k concepts)
Σ: *k x k* diagonal matrix (strength of each 'concept')
V: k

B. Principal Component Analysis(PCA*)*

PCA is an analysis tool for identifying patterns in data and expressing these data in such a way that it highlights their similarities and differences[17].

PCA is unsupervised algorithm as follows:
Algorithm:

1. Organise data into n*m matrixwhere m is measurement type and n is number of samples[15].
2. Subtract off mean from each measurement type[15].
3. Calculate Covariance Matrix[15].
4. Calculate Eigen Values and Eigen Vectors from the Covariance Matrix[15].

Module 3 : Applying Clustering Approaches and IR

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters) [6]. Clustering is a data mining (machine learning) technique used to place data elements into related groups without advance knowledge of the group definitions[15].

A. K-Means Clustering

K-means is the most traditional and simple approach for data clustering. This approach clusters the all input documents into specified and user pre-defined value K. Here in this research we have clustered the entire dataset into four clusters. For effective clustering the Euclidean formula is made use of for calculating distance between data instances. Euclidean Distance Formula for K-means implementation is-

$$\mathbf{J(V) = \sum_{i=1}^{c} \sum_{j=1}^{ci} (\| xi - vj \|)2}$$

where,

'*|xi - vj|*' is the Euclidean distance between *xi* and *vj*.
'ci' is the number of data points in i^{th} cluster.
'c' is the number of cluster centers.
'**xi**is the data points in i^{th} cluster.
'**vj**' is the center of j^{th} cluster.

Steps:

1. Assign first k data instances as K clusters i.e. cluster centroids.
2. For every incoming data instance till the end of data set check Euclidean distance from all K cluster centroids, assign the cluster with minimum distance.
3. Repeat step2 until well densed and properly separated clusters are formed .

B. Fuzzy Clustering (FCM)

Fuzzy clustering is also known as Fuzzy C-means. This algorithm assigns every data instance a meaning full cluster. The clustering is based upon membership function. Each and every data instance in the dataset has some value of membership function i.e. true or false[18]. Based on this value the cluster is assigned to that data member. FCM is an variation to hard clustering approach[18]. Thus the clusters are not well separated and the algorithm does not follow the clustering hypothesis.

C. Hierarchical Clustering

This is a tree based approach, herein two clusters are grouped together to a root cluster. A threshold value is predefined such as two clusters cannot come together in one root if the distance between them is more than the threshold value.

The Hierarchical algorithm works in following way:

1. Each data point x1,x2,…..,xn initially resides in its own individual cluster C1,C2,….,Cn.
2. The distance between two clusters is calculated and compared with the threshold.
3. Nearest clusters following the threshold constraints are merged to form a single cluster.
4. The process is repeated till single cluster is formed. This approach ultimately produces a cluster tree with leaf nodes as individual data instances from the input data set.

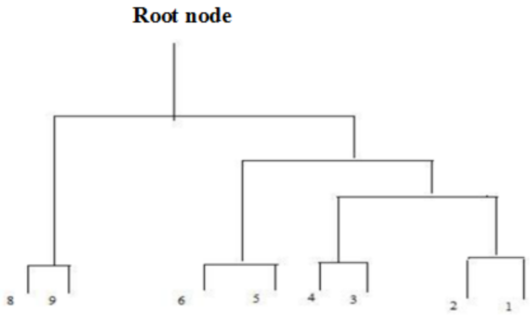


Fig.2. Hierarchical Clustering Tree

III. EXPERIMENTAL RESULTS

C50test dataset was used for performing all the experiments[14]. It contains 2500 files which were preprocessed using dimensionality reduction techniques

IJARCCE

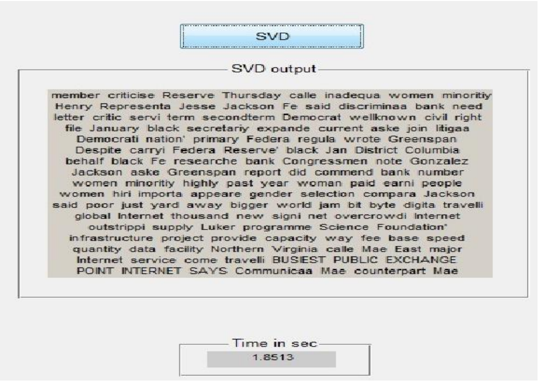
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 3, March 2016

like SVD and PCA. Dimensionally reduced word matrix was then clustered using Clustering technique like K-means, FCM and Hierarchical Clustering. All the obtained clusters were given as input to the Information retrieval system.

SVD Matrix with Computation time

The dataset obtained after preprocessing is treated with SVD algorithm to produce a dimensionality reduced word matrix as shown in Fig.4, the time required for this computation was 1.8513 seconds.



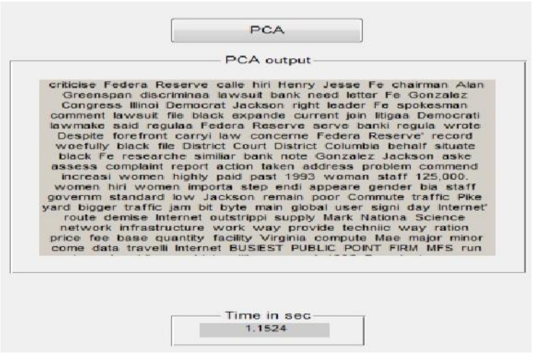
member criticise Reserve Thursday calle inadequa women minoritiy Henry Representa Jesse Jackson Fe said discriminaa bank need letter critic servi term secondterm Democrat wellknown civil right file January black secretariy expande current aske join litigaa Democrati nation' primary Federa regula wrote Greenspan Despite carryi Federa Reserve' black Jan District Columbia behalf black Fe researche bank Congressmen note Gonzalez Jackson aske Greenspan report did commend bank number women minoritiy highly past year woman paid earni people women hiri importa appeare gender selection compara Jackson said poor just yard away bigger world jam bit byte digita travelli global Internet thousand new signi net overcrowdi Internet outstrippi supply Luker programme Science Foundation' infrastructure project provide capacity way fee base speed quantity data facility Northern Virginia calle Mae East major Internet service come travelli BUSIEST PUBLIC EXCHANGE POINT INTERNET SAYS Communicaa Mae counterpart Mae

1.8513

Fig.3.SVD Computation

PCA Matrix with Computation time

The dataset obtained after preprocessing is treated with PCA algorithm to produce a dimensionality reduced word matrix as shown in Fig.5, the time required for this computation was 1.1524 seconds.



criticise Federa Reserve calle hiri Henry Jesse Fe chairman Alan Greenspan discriminaa lawsuit bank need letter Fe Gonzalez Congress Illinoi Democrat Jackson right leader Fe spokesman comment lawsuit file black expande current join litigaa Democrati lawmake said regulaa Federa Reserve serve banki regula wrote Despite forefront carryi law concerne Federa Reserve' record woefully black file District Court District Columbia behalf situate black Fe researche similiar bank note Gonzalez Jackson aske assess complaint report action taken address problem commend increasi women highly paid past 1993 woman staff 125,000. women hiri women importa step endi appeare gender bia staff governm standard low Jackson remain poor Commute traffic Pike yard bigger traffic jam bit byte main global user signi day Internet' route demise Internet outstrippi supply Mark Nationa Science network infrastructure work way provide techniic way ration price fee base quantity facility Virginia compute Mae major minor come data travelli Internet BUSIEST PUBLIC POINT FIRM MFS run

1.1524

Fig.4.PCA Computation

Clusters formed using k-means technique on SVD results

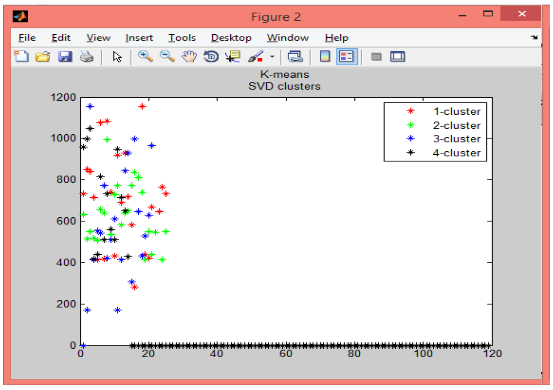


Fig.5.K-means-SVD Data Clusters

Clusters formed using k-means technique on PCA results

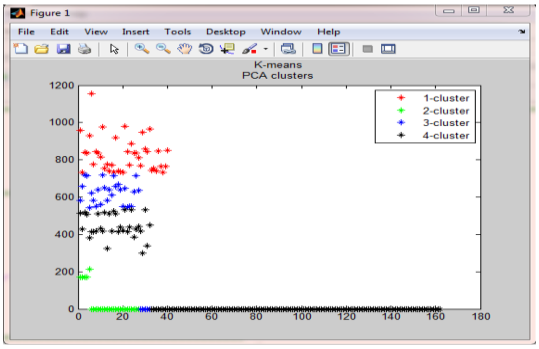


Fig.6.K-means-PCA Data Clusters

Clusters formed using fuzzy technique on SVD results

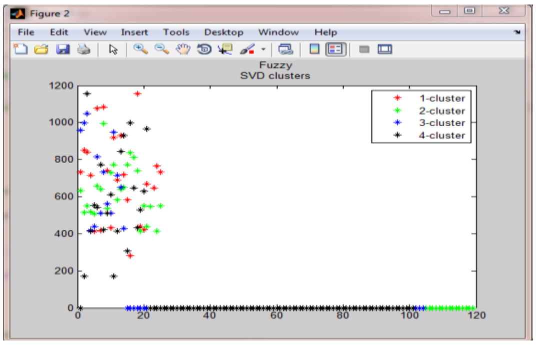


Fig.7. Fuzzy-SVD Data Clusters

Clusters formed using fuzzy technique on PCA results

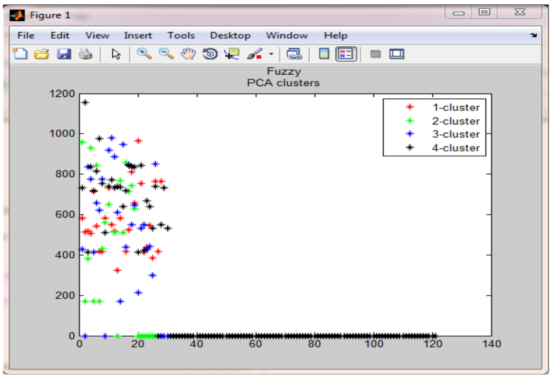


Fig.8. Fuzzy-PCA Data Clusters

Clusters formed using hierarchical technique on SVD results

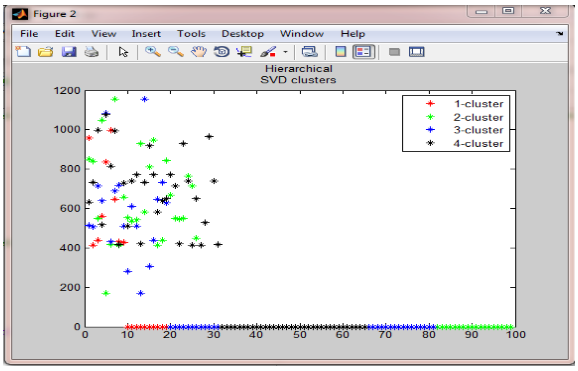


Fig.9. Hierarchical-SVD Data Clusters

DOI 10.17148/IJARCCE.2016.5381

Clusters formed using hierarchical technique on PCA results

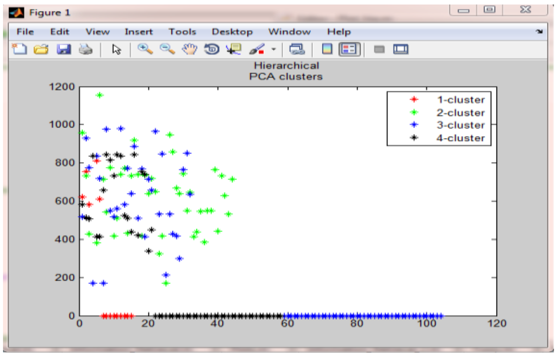
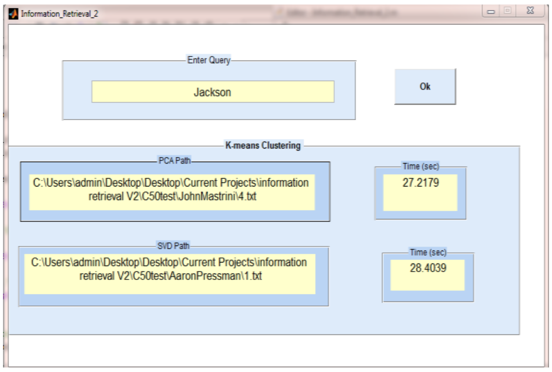


Fig.10. Hierarchical-PCA Data Clusters

Information Retrieval using K-means Clusters



Jackson

C:\Users\admin\Desktop\Desktop\Current Projects\information retrieval V2\C50test\JohnMastrini\4.txt

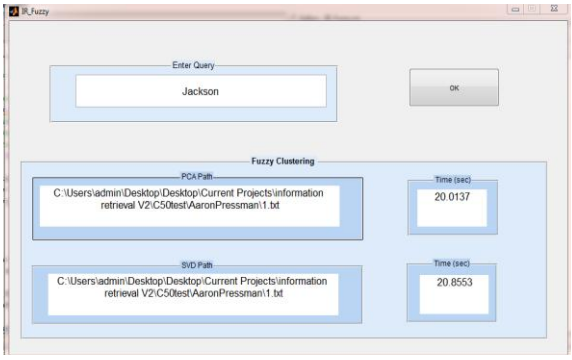
27.2179

C:\Users\admin\Desktop\Desktop\Current Projects\information retrieval V2\C50test\AaronPressman\1.txt

28.4039

Fig.11. K-means IR

Information Retrieval using Fuzzy Clusters



Jackson

C:\Users\admin\Desktop\Desktop\Current Projects\information retrieval V2\C50test\AaronPressman\1.txt

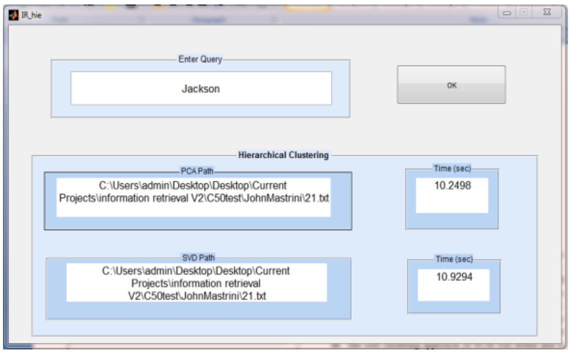
20.0137

C:\Users\admin\Desktop\Desktop\Current Projects\information retrieval V2\C50test\AaronPressman\1.txt

20.8553

Fig.12. Fuzzy Clustering IR

Information Retrieval using Hierarchical Clusters



Jackson

C:\Users\admin\Desktop\Desktop\Current Projects\information retrieval V2\C50test\JohnMastrini\21.txt

10.2498

C:\Users\admin\Desktop\Desktop\Current Projects\information retrieval V2\C50test\JohnMastrini\21.txt

10.9294

Fig.13. Hierarchical Clustering IR

IV. EXPERIMENTAL ANALYSIS

In this research C50 dataset was examined under various possibilities for making it effective for information retrieval. Initially the DR techniques like SVD and PCA were compared to notice that PCA was far better in any circumstances. Later Various clustering approaches were prominently compared. After analyzing it was found that though k-means cluster formation was better but it failed to give optimum results at IR. The soft clustering approach of FCM was tested and it gave good results for PCA dataset for IR. Hierarchical clustering was the lastly examined approach which yielded excellent retrieval results based on time and reliability.

TABLE 1: SAMPLING OF DATASET QUERIES FOR VARIOUS CLUSTERING TECHNIQUES

Query	Retrieval Time		
	K-means	Fuzzy	Hierarchical
Millennium	34.83	20.22	10.22
Congress	31.47	21.04	9.34
addition	32.48	20.88	10.79
Broke	33.80	22.20	10.68
Japan	33.34	21.79	8.92
China	25.42	20.22	11.33
Tuesday	34.12	19.33	10.82
Party	24.41	17.55	10.78
lawsuit	35.07	18.03	12.33
Week	30.66	22.11	14.54
Authority	31.22	21.08	13.9
Department	30.05	15.99	12.57
director	29.88	20.80	11.45
speaking	26.77	18.45	12.54
Aviation	28.4	20.15	11.7

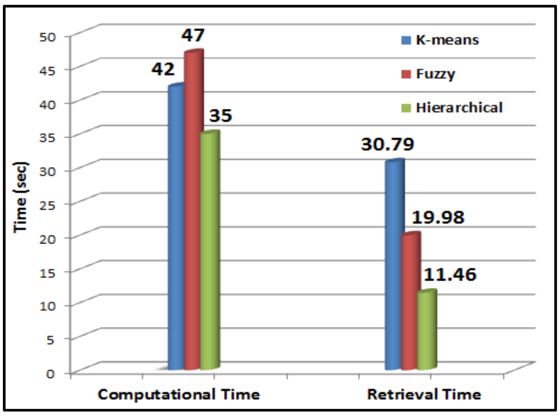


Fig.14. Query retrieval time pattern for various techniques

TABLE 2: COMPARING CLUSTERING APPROACHES

Parameters	K-means	Fuzzy	Hierarchical
Computational time (sec)	42	47	35
Max. Retrieval time for any query (sec)	30.79	19.98	11.46
Cluster Efficiency for IR	Good	Better	Excellent

DOI 10.17148/IJARCCE.2016.5381

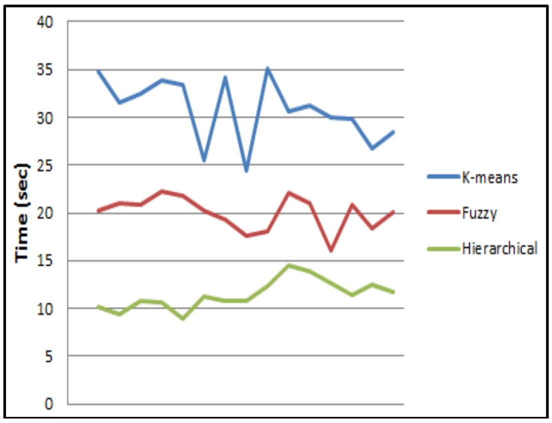


Fig.15. Time based comparison of clustering

V. CONCLUSION

This research ultimately contributes to the comprehensive choices of techniques and algorithms to be used for Information Retrieval System. A totally new approach to the development of IR has been laid forward by comparing various techniques. A brief comparison of DR techniques like SVD and PCA proves that PCA can be better for effective and promising results. At the later stage Clustering techniques of totally different methodology and mechanism were examined with same data set; the very next output of clusters was treated for IR system to get better conclusions over the techniques and variety of approach combinations carried out throughout the study.

All the clustering approaches like K-means clustering, Fuzzy clustering and the Hierarchical clustering gave different output for cluster plots, time required and the retrieval time of query in IR system. Based on all the results put forward in above experimental results, it appeared to be very distinct then other IR developing approaches and also yielded such relevant and better results than traditional systems. The results to be focused on were one which were obtained from Hierarchical clustering on the PCA dataset. These results were excellent and this methodology of IR can prove boon to the high dimensional dataset managing machines or systems. Thus, the overall study led to a better, effective, efficient, reliable, relevant and excellent Information System which can be user friendly and applied anywhere on textual datasets for ease of data handling, management and access through retrieval.In near future same system will be examined with various datasets for more confident solutions and conclusions.

REFERENCES

1. V. Srividhya, R. Anitha , " Evaluating Preprocessing Techniques in Text Categorization ",ISSN 0974-0767,International Journal of Computer Science and Application Issue 2010
2. Nguyen Hung Son, "Data Cleaning and Data Preprocessing".
3. Lei Yu Binghamton University, JiepingYe, Huan Liu ,Arizona State University, “Dimensionality Reduction for data mining-Techniques, Applications and Trends”.
4. Ch. Aswani Kumar, "Analysis of Unsupervised Dimensionality Reduction Techniques", Com SIS Vol. 6, No. 2, December 2009.
5. C.Ramasubramanian, R.Ramya, "Effective Pre-Processing Activities in Text Mining using Improved Porter’s Stemming Algorithm", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 12, December 2013.
6. Rui Tang, Simon Fong, Xin-She Yang, Suash Deb,” Integrating nature-inspired optimization algorithms to k-means clustering”, 978-1-4673-2430-4/12/$31.00 ©2012 IEEE.
7. Carlos Cobos, Henry Muñoz-Collazos, RicharUrbano-Muñoz, Martha Mendoza, Elizabeth Leónc, Enrique Herrera-Viedma “Clustering Of Web Search Results Based On The Cuckoo Search Algorithm And Balanced Bayesian Information Criterion ” ELSEVIER Publication, 2014 Elsevier Inc. All rights reserved ,21 May 2014
8. Agnihotri, D.; Verma, K.; Tripathi, P., "Pattern and Cluster Mining on Text Data," Communication Systems and Network Technologies (CSNT), 2014 Fourth International Conference on, vol., no., pp.428,432, 7-9 April 2014
9. Patil, L.H.; Atique, M., "A novel approach for feature selection method TF-IDF in document clustering," Advance Computing Conference (IACC), 2013 IEEE 3rd International, vol., no., pp.858,862, 22-23 Feb. 2013
10. RasmusElsborg Madsen, Lars Kai Hansen and Ole Winther, "Singular Value Decomposition and Principal Component Analysis",February 2004.
11. https://www.irisa.fr/sage/bernard/publis/SVD-Chapter06.pdf
12. https://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Dimensionality_Reduction/Singular_Value_Decomposition
13. https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm
14. http://archive.ics.uci.edu/ml/datasets/Reuter_50_50
15. Poonam P. Rajurkar, Aditya G. Bhor, Komal K. Rahane, Neha S. Pathak,"Efficient Information Retrieval through Comparison of Dimensionality Reduction Techniques with Clustering Approach ", International Journal of Computer Applications (0975 – 8887), Volume 129 – No.4, November2015.
16. Komal Rahane, Poonam Rajurkar, Neha Pathak, Aditya Bhor, Asst. Prof. Anagha N. Chaudhari," Comprehensive Comparison of Various Approaches for Implementation of Expert IR System through Pre-processing and Clustering",International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 3 Issue: 10 5752 - 5757, October 2015.
17. Neha S. Pathak, Poonam P.Rajurkar, Aditya G. Bhor, Komal K. Rahane, Anagha N. Chaudhari," EFFECTIVE APPROACH TOWARDS EXPERT IR SYSTEM THROUGH COMPARISON OF VARIOUS PREPROCESSING TECHNIQUES", International Journal of Advances in Engineering Science and Technology, Volume 4, Number 3 ISSN: 2319-1120 /V4N3: 118-123 © IJAEST,2015.
18. Anagha Chaudhari,” A Novel Approach for Development of an Expert IR System Using Dimensionality Reduction Techniques and Clustering Approaches for High Dimensionality Dataset”, International Journal of Computer Applications Volume 128 – No.2, October 2015

DOI 10.17148/IJARCCE.2016.5381