A Novel Approach for Development of an Expert IR System using Dimensionality Reduction Techniques and Clustering Approaches for High Dimensionality Dataset

Anagha N. Chaudhari
Assistant Professor, Department of Information Technology
Pimpri Chinchwad College of Engineering
Pune, India

ABSTRACT

In day to day life huge amount of electronic data is generated from various resources. Such data is literally large and not easy to work with for storage and retrieval. This type of data can be treated with various efficient techniques for cleaning, compression and sorting of data. Preprocessing 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. This data later can be clustered for better information retrieval. 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 Method.

1. INTRODUCTION

For the complex data sets there is a problem in retrieval of the necessary information from particular records. As the original datasets are multidimensional in nature, so for retrieving the specific information, datasets need to be reduced in dimensionality. Hence, for this there are different dimensionality reduction techniques and by using these techniques the datasets are first reduced and further clustered using efficient clustering algorithm. Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) are used for dimensionality reduction and further obtained outputs from both are applied with the K-means clustering.

Modules

Module 1: Preprocessing

Module 2: Application of Dimensionality Reduction Techniques

Module 3: Applying Clustering Approaches

2. SYSTEM ARCHITECTURE

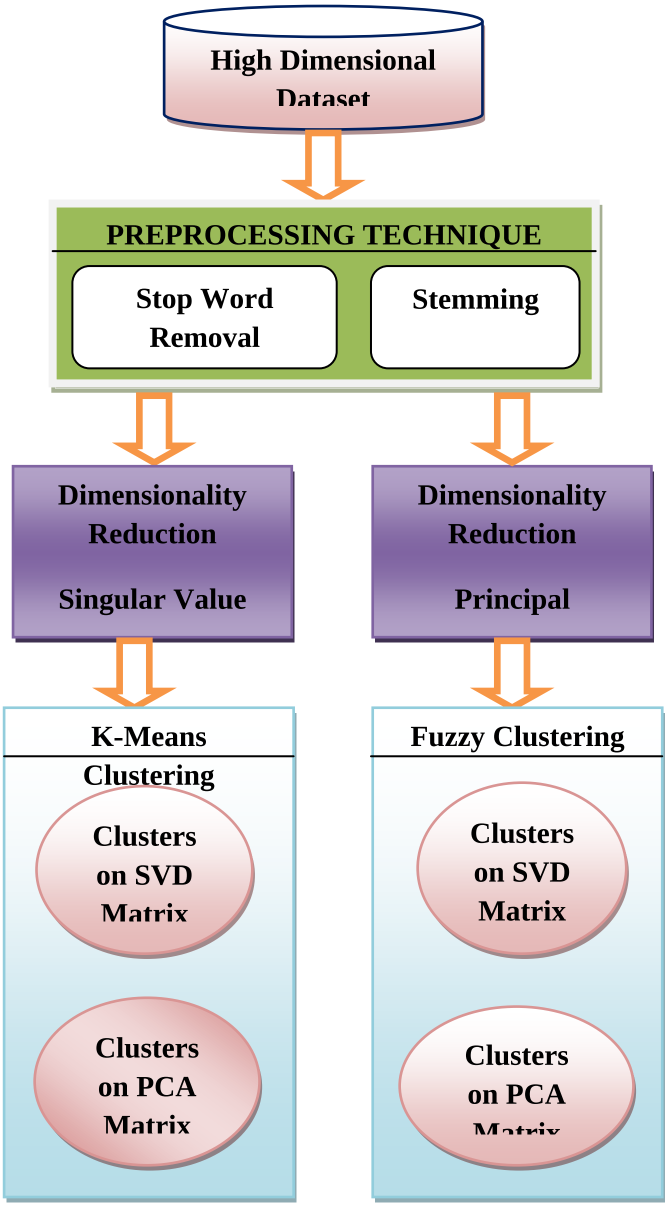


Figure 1 System Architecture

Figure 1 represents the detailed system architecture as explained in further sections.

Module 1: Preprocessing

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

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. 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].The data preprocessing techniques used in this paper are stop word removal and stemming.”. 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].

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 the computational efficiency and accuracy in the analysis of data. The problem of dimension reduction can be defined mathematically as follows : given a *r*-dimensional random vector **Y**=(*y*1,*y*2,….*yr*)T, it’s main objective is to find a representation of lower dimension **P**=(*p*1,*p*2,…,*pk*)T, where *k*<*r*, which preserves the content of the original data, according to some criteria[4].

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).

Dimensionality Reduction is done when:

- Irrelevant features exist in data.
- High dimensional data visualization.

1. Singular Value Decomposition(SVD)

In data mining, this algorithm can be used to better understand a database by showing the number of important dimensions and also to simplify it, by reducing of the number of attributes that are used in a data mining process [12]. This reduction removes unnecessary data that are linearly dependent in the point of view of Linear Algebra[12].In computational science, it is commonly applied in Information Retrieval (IR)[11].SVD can be implemented using formula shown in Figure 2.

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

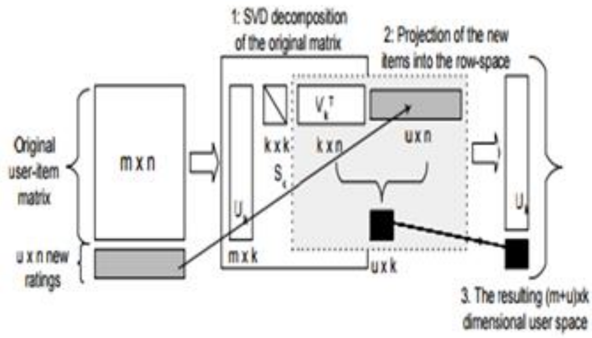

1: SVD decomposition of the original matrix
2: Projection of the new items into the row-space
Original user-item matrix
m x n
u x n new ratings
U_k
m x k
k x k
S_k
V_k^T
k x n
u x n
u x k
U_k
3. The resulting (m+u)xk dimensional user space

Figure 2 Schematic diagram of SVD

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

2. PRINCIPAL COMPONENT ANALYSIS(PCA*)*

In principal component analysis we find the directions in the data with the most variation, i.e. the eigenvectors corresponding to the largest eigen values of the covariance matrix, and project the data onto these directions [10].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. PCA is unsupervised algorithm. PCA ignores the class labels in the datasets. This algorithm is used to find the direction that maximizes the variance in the datasets.

Algorithm:

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

Module 3 : Applying Clustering Approaches

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. The most popular clustering technique is k-means clustering. K-means clustering is a data mining/machine learning algorithm used to cluster observations into groups of related observations without any prior knowledge of those relationships [7]. The k-means is one of the simplest clustering techniques and it is commonly used in data mining, biometrics and related fields [8].Euclidean Distance Formula for K-means implementation is-

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} (\|x_i - v_j\|)^2$$

where,

'$|x_i - v_j|$' is the Euclidean distance between x_i and v_j.

'c_i' is the number of data points in i^{th} cluster.

'c' is the number of cluster centers.

'x_i is the data points in i^{th} cluster.

'v_j' is the center of j^{th} cluster.

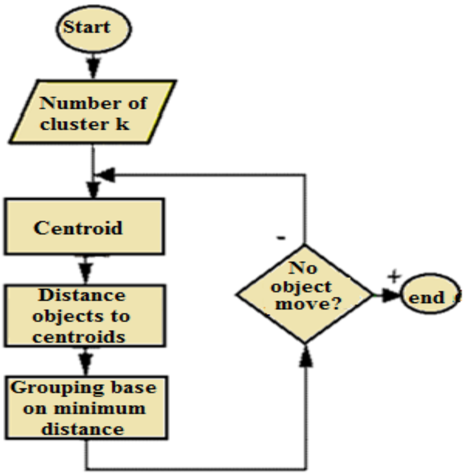

Start
Number of cluster k
Centroid
Distance objects to centroids
Grouping base on minimum distance
No object move?
-
+
end

Figure 3 K-means Flowchart

K-means algorithm

Let X = {x1,x2,x3,……..,xn} be the set of data points and

V = {v1,v2,…….,vc} be the set of centers.

1. Randomly select *'c'* cluster centres [13].
2. Calculate the distance between each data point and cluster centre[13].
3. Assign the data point to the cluster centre whose distance from the cluster centre is minimum of all the cluster centres [13].
4. Recalculate the new cluster centre using:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

   where, *'ci'* represents the number of data points in *ith* cluster[13].

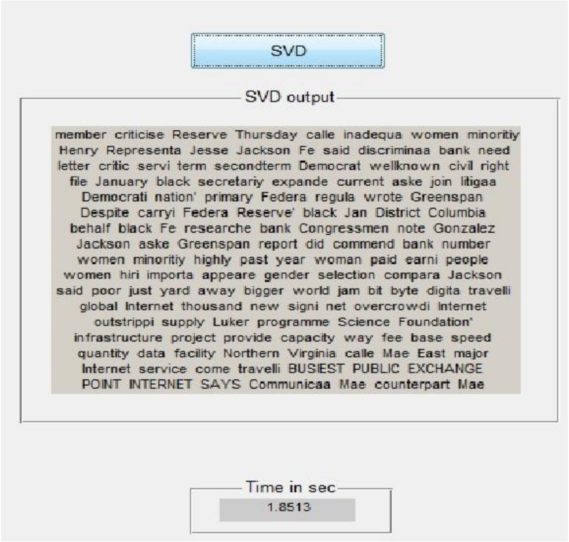
5. Recalculate the distance between each data point and new obtained cluster centres [13].
6. If no data point was reassigned then stop, otherwise repeat from step 3[13] as shown in Figure 3.

3. EXPERIMENTAL RESULTS

C50test dataset was used for performing all the experiments [14]. It contains 2500 files which were preprocessed using dimensionality reduction techniques like SVD and PCA. Dimensionally reduced word matrix was then clustered using Clustering technique like K-means.

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 Figure 4, the time required for this computation was 1.8513 seconds.



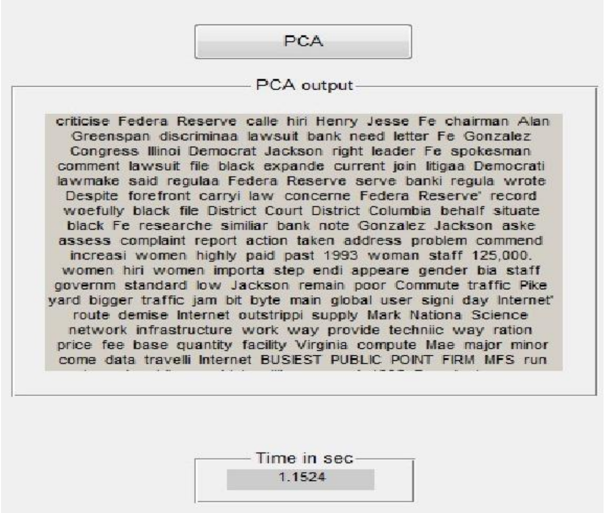
```
member criticise Reserve Thursday calle inadequa women minoritiy
Henry Representa Jesse Jackson Fe said discriminaa bank need
letter critic servi term secondterm Democrat wellknown civil right
file January black secretariy expande current aske join litigaa
Democrati nation' primary Federa regula 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
```

Figure 4 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 Figure 5, the time required for this computation was 1.1524 seconds.



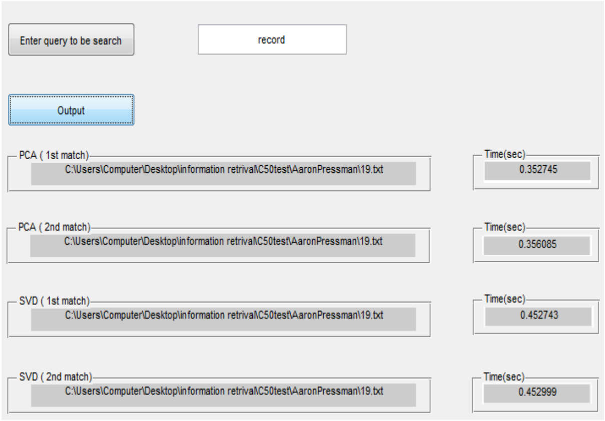
```
criticise Federa Reserve calle hiri Henry Jesse Fe chairman Alan
Greenspan discriminaa lawsuit bank need letter Fe Gonzalez
Congress Illinoi Democrat Jackson right leader Fe spokesman
comment lawsuit file black expande current join litigaa Democrati
lawmake said regulaa Federa Reserve serve banki 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
```

Figure 5 PCA Computation

Retreival of word from SVD and PCA matrices

The dimensionally reduced word matrices obtained by SVD and PCA were used for retreiving words from it through an user interface by entering user desired query (Figure 6). The retreival time for matching user query and displaying its path in dataset documents was less for PCA, proving it better than SVD.



record

PCA (1st match): C:\Users\Computer\Desktop\information retrival\C50test\AaronPressman\19.txt — Time(sec): 0.352745

PCA (2nd match): C:\Users\Computer\Desktop\information retrival\C50test\AaronPressman\19.txt — Time(sec): 0.356085

SVD (1st match): C:\Users\Computer\Desktop\information retrival\C50test\AaronPressman\19.txt — Time(sec): 0.452743

SVD (2nd match): C:\Users\Computer\Desktop\information retrival\C50test\AaronPressman\19.txt — Time(sec): 0.452999

Figure 6 Query Retrieval

Clusters formed using k-means technique on SVD results

The output in the form of dimensionally reduced word matrix by applying SVD to the preprocessed dataset was clustered using K-means clustering, by considering k=4, i.e 4 clusters were formed (Figure 7).

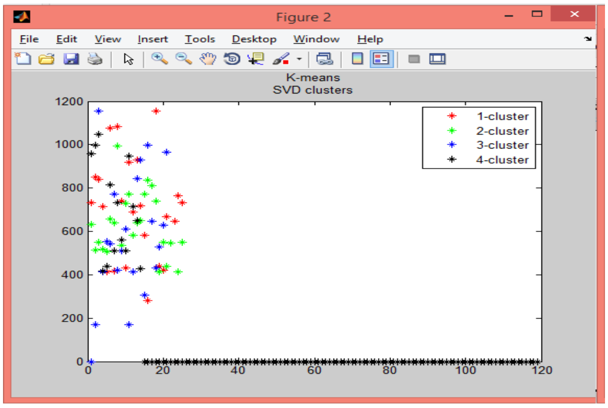


Figure 7 SVD Data Clusters

Clusters formed using k-means technique on PCA results

Dimensionally reduced word matrix obtained by applying PCA on preprocessed dataset was clustered using K-means clustering forming 4 clusters which were prominently categorized (Figure 8).

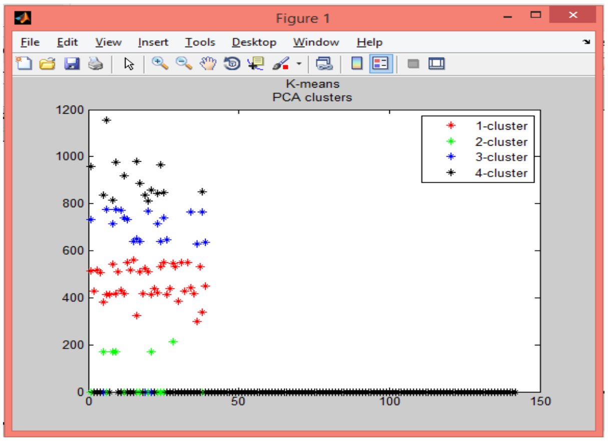


Figure 8 PCA Data Clusters

4. EXPERIMENTAL ANALYSIS (SVD V/s PCA)

Both the techniques were applied to pre-processed C50test dataset. The observations made depict that PCA being the next version of SVD proves to be better in many ways, i.e. the computational time of PCA technique for formation of dimensionally reduced weight matrix was less than that of SVD as well as the retrieval time for retrieving any specific word in the dimensionally reduced word matrix through user query was significantly less for PCA algorithm (Figure 9).And further when computations of PCA and SVD were clustered using K-means Clustering algorithm, the clusters formed on SVD data were scattered and not prominent where as those formed on PCA data were accurate, categorized and better than SVD.

Table 1. Comparison of SVD and PCA Based on Various Experimental Parameters

Parameters	SVD	PCA
Computational time (sec)	1.8513	1.1524
Retrieval time (sec)	0.4529	0.3527
Reduced Matrix size (doc x max_terms)	2500 x 289	2500 x 290
Clusters Formed	Scattered	Categorized and Accurate

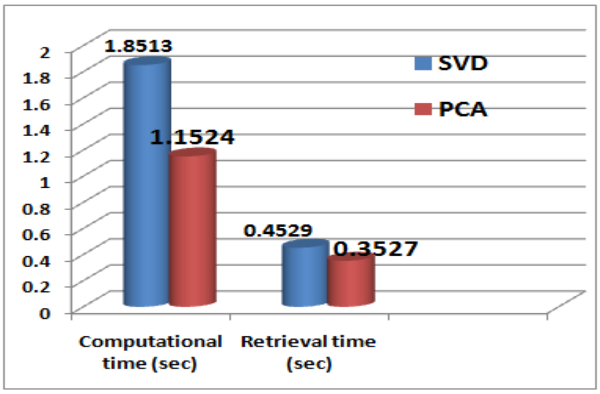


Figure 9 SVD and PCA comparison graph

5. EXPERIMENTAL ANALYSIS (K-Means V/s Fuzzy Clustering)

Fuzzy logic is a mathematical logic model in which truth can be partial i.e. it can have value between 0 and 1, that is completely false and completely true [16]. It is based on approximate reasoning instead of exact reasoning. Using Fuzzy Logic and text-mining we can cluster similar documents together. Document Clustering is used by the computer to group documents into meaningful groups. We have used c-means algorithm. This algorithm is classified into two types- Hard c-means and FCM [16].

Hard c-means algorithm is used to cluster 'm' observations into 'c' clusters. Each cluster has its cluster centre. The observation belongs to the cluster which has the least distance from it. The clustering is crisp, which means that each observation is clustered into one and only one cluster.

FCM is a variation of the hard c-means clustering algorithm. Every observation here has a membership value associated with each of the clusters which is related inversely to the distance of that observation from the centre of the cluster [15].The output in the form of dimensionally reduced word matrix by applying SVD to the preprocessed dataset was clustered using Fuzzy clustering, by considering k=4, i.e 4 clusters were formed (Figure 10).

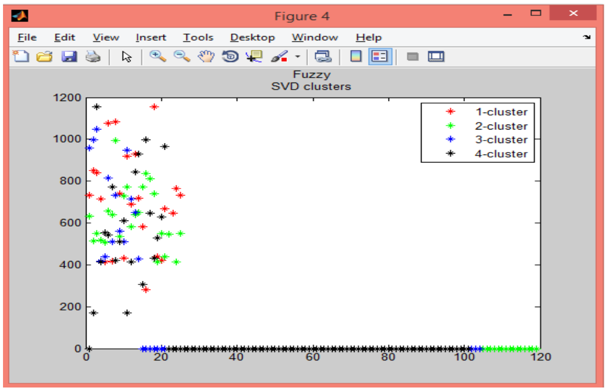


Figure 10 SVD-Fuzzy Clusters in Matlab 2014

The output in the form of dimensionally reduced word matrix by applying PCA to the preprocessed dataset was clustered using Fuzzy clustering, by considering k=4, i.e. 4 clusters were formed(Figure 11).

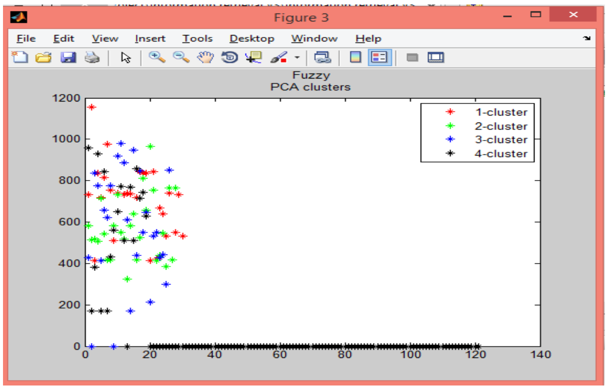
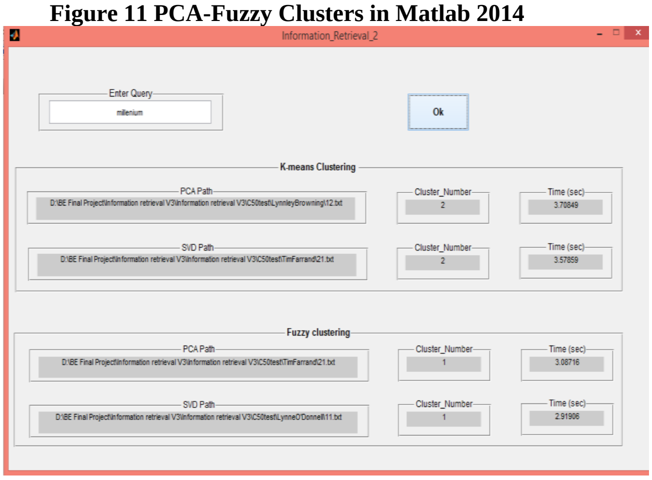


Figure 11 PCA-Fuzzy Clusters in Matlab 2014



Enter Query: milenium

K-means Clustering

PCA Path: D:\BE Final Project\Information retrieval V3\Information retrieval V3\C50test\LynnleyBrowning\12.txt
Cluster_Number: 2
Time (sec): 3.70849

SVD Path: D:\BE Final Project\Information retrieval V3\Information retrieval V3\C50test\TimFarrand\21.txt
Cluster_Number: 2
Time (sec): 3.57859

Fuzzy clustering

PCA Path: D:\BE Final Project\Information retrieval V3\Information retrieval V3\C50test\TimFarrand\21.txt
Cluster_Number: 1
Time (sec): 3.08716

SVD Path: D:\BE Final Project\Information retrieval V3\Information retrieval V3\C50test\LynneO'Donnell\11.txt
Cluster_Number: 1
Time (sec): 2.91906

Figure 12 Information Retrieval from K-means and Fuzzy Clusters

In above figure (Figure 12), IR system of K-means and Fuzzywith SVD & PCA approaches is represented where it is clearly visible that for a sample word "milllenium" the retrieval time (in ms) of Fuzzy clustering with SVD (2.9 ms) & PCA(3.08 ms) is much less than K-means clustering with SVD (3.5 ms) & PCA (3.7 ms).

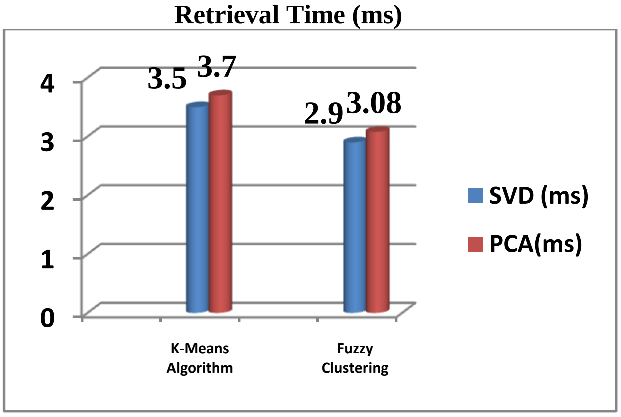


Figure 13 Comparison of Retrieval Time of K-means and Fuzzy Clusters

Figure 13 depicts the graphical representation of retrieval time comparison of K-Means and Fuzzy Clustering with SVD & PCA approaches. It is proved that FCM takes less retrieval time with SVD & PCA computations when compared with K-Means clustering algorithm.

6. CONCLUSION

This paper mainly contributes to provide a new approach with comparative study of the dimensionality reduction techniques as SVD and PCA to improve the performance of clustering. Users can make comprehensive choices among various available dimensionality reduction techniques referencing this study. The main objective is to achieve best performance of K-means clustering by treating original dataset like C50test with pre-processing techniques like stop-word removal followed by stemming and then later with dimensionality reduction techniques. Hence after obtaining SVD-PCA computations as SVD-K-means, PCA-K-means; it proves that K-means done on data computed by PCA technique gives accurate, categorized and better results than SVD. This research has shown comparable results of available techniques on C50test dataset and a better approach for relevant Information Retrieval system. Also, results focuses on the better performance of Fuzzy clustering over K- Means algorithm as retrieval time of FCM is much lesser than K-Means approach with SVD and PCA computations, thus, resulting into an expert IR system.

7. ACKNOWLEDGEMENT

This work has received research grant and is funded by **BCUD** under **Savitribai Phule Pune University, Pune**. Many thanks to BCUD for giving me this opportunity to carry out the research work in the field of Data Mining and Information Retrieval.

8. 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 datamining-Techniques, Applications and Trends".

[4] Ch. Aswani Kumar, "Analysis of Unsupervised Dimensionality Reduction Techniques", ComSIS 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 EngineeringVol. 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 andPrincipal 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] Sumit Goswami, Mayank Singh Shishodia; "A fuzzy based approach to stylometric analysis of blogger‟s age and gender"; HIS 2012: 47-5

[16] Ross, T. J. (2010); "Fuzzy Logic with Engineering Applications", Third Edition, John Wiley & Sons, Ltd, Chichester, UK.