International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 3 Issue: 10 5752 - 5757

Comprehensive Comparison of Various Approaches for Implementation of Expert IR System through Pre-processing and Clustering

Ms. Komal Rahane, Ms. Poonam Rajurkar, Ms. Neha Pathak, Mr.Aditya Bhor
Asst. Prof. Anagha N. Chaudhari
Department of Information Technology
Pimpri Chinchwad College Of Engineering
Pune,India.

***Abstract*—** We live in a digital world wherein tremendous electronic data evolves out every day, generating from huge amount of sources present. This data is tedious and nearly impossible to manage as it being literally large. Data storage and retrieval becomes a truly difficult task. Thus through data mining approach 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, using tool based as well as programming based approach for providing comprehensive choices among the users.

***Keywords* —** *High Dimensional Datasets, Dimensionality reduction, SVD, PCA, Clustering, K-means*

*****

I. INTRODUCTION

In this really fascinating world, digital dimensions have gained a significant importance. Thus, the data produced and datasets formed out of it a high dimensional datasets. It becomes hard to deal with such data while desired and relevant information needs to be retrieved with accuracy as well as less possible time. Thus, dealing with this type of datasets need dimensionality reduction. 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.This whole system can be implemented using tool based as well as programming based approach thus using MATLAB-R2104a as a tool and JAVA programming language for the programming based approach it can be significantly explored about the better way by observing the results in different aspects.

II. MODULES

Module 1: Preprocessing.
Module 2: Dimensionality Reduction Techniques.
Module 3: Information Retrieval Through Clustering.

A. Module1: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].

III. SYSTEM ARCHITECTURE

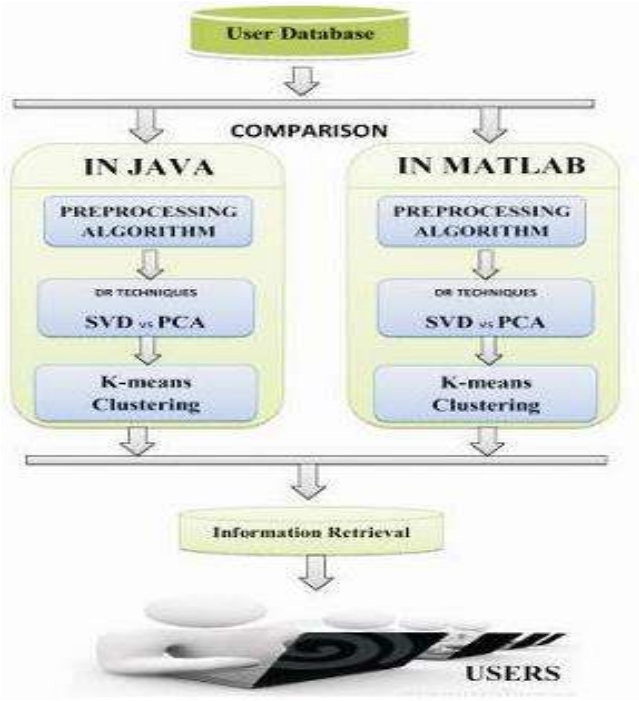

Fig. 1 System Architecture

Algorithm: Preprocessing

Input: Dataset (Contains n no. of text files)
Step 1: for each document in Dataset remove stop-words.
Step 2: for all similar words produce a root
and do stemming
Step 3: produce a document * word matrix for
further processing
Output: Preprocessed dataset.

B. Module2 - 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 $\mathbf{Y}=(y1,y2,\ldots yr)$T, it's main objective is to find a representation of lower dimension $\mathbf{P}=(p1,p2,\ldots,pk)$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).

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 Fig.2.

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

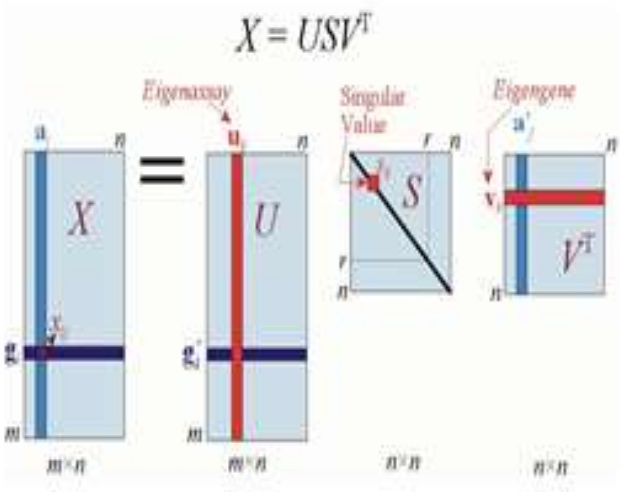

Fig. 2 Schematic diagram of SVD

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

C. Module3 - 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].

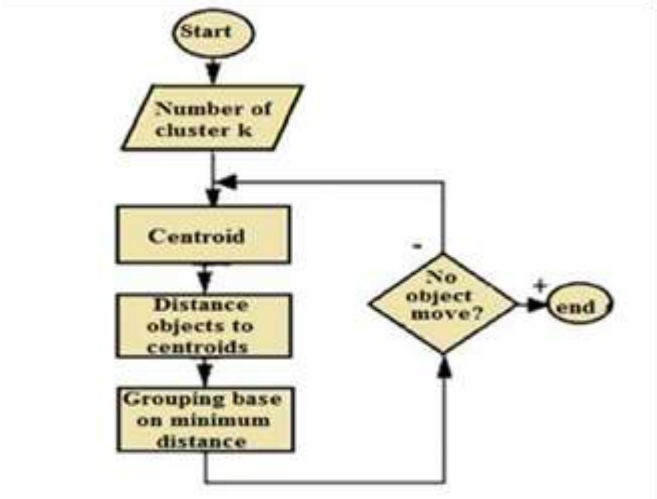

Fig. 3 K-means Flowchart

IV. 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.

A. Preprocessed Dataset

The original dataset C50 test is treated with preprocessing, all the basic English stopwords also called as fuzzy words are eliminated and also the similar words are stemmed accordingly, resultant matrix is formed as a matrix with numerical values obtained in accordance with the ASCII values of the terms, hence a numerical preprocessed matrix is generated which can be easily used for used for further processing.

1) Preprocessed Data matrix obtained in Java

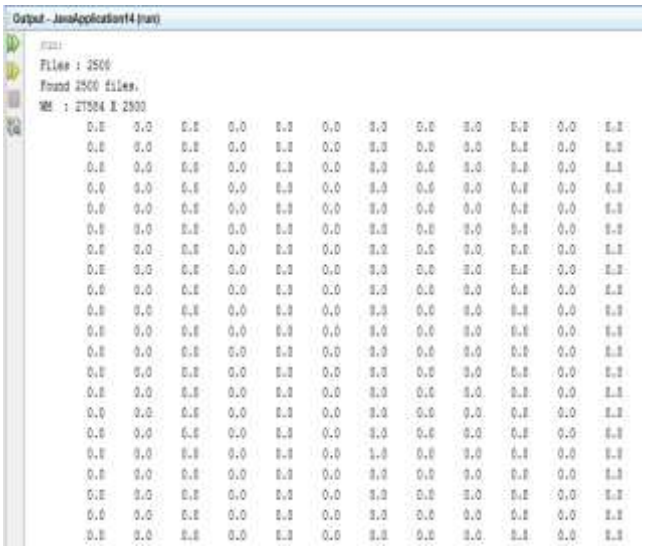

Fig. 4 Stopword List

2) Preprocessed Data matrix obtained in MATLAB

	1	2	3	4	5	6	7	8	9	10
1	314	632	836	959	583	732	852	513	887	840
2	730	554	735	825	393	848	825	454	432	434
3	520	962	855	446	716	393	735	536	542	641
4	841	330	769	420	528	323	431	417	841	529
5	802	869	613	932	954	553	637	214	825	656
6	615	737	852	435	412	513	557	635	410	898
7	836	657	613	1099	674	656	428	614	472	672
8	836	657	613	1099	674	656	428	614	472	672
9	583	412	737	529	642	214	825	1226	751	865
10	734	622	541	932	727	1549	653	772	648	1311
11	649	412	622	1099	337	647	648	725	635	440
12	649	412	622	1099	337	647	648	725	635	440
13	757	661	644	1198	410	451	981	766	737	214
14	330	534	825	635	411	635	441	869	447	1113
15	642	595	546	298	421	404	394	520	426	519
16	642	595	546	298	421	404	394	520	426	519
17	214	766	939	643	870	765	677	875	970	730

Fig. 5 Document * term Matrix

B. SVD and PCA computations on Preprocessed data matrix

The dataset obtained after preprocessing is treated with both SVD and PCA techniques, for extraction of only relevant and important terms from the riginal dataset, hene reducing its dimensionality. The implementation is done using the formulae as explained previously in section II.

1) Java computations (SVD and PCA)

a. SVD Matrix with Computation time in JAVA

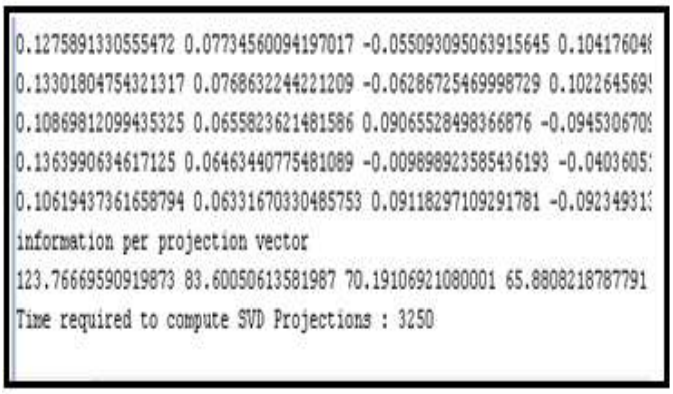

Fig. 6. SVD Results in Java

b. PCA Matrix with Computation time in JAVA

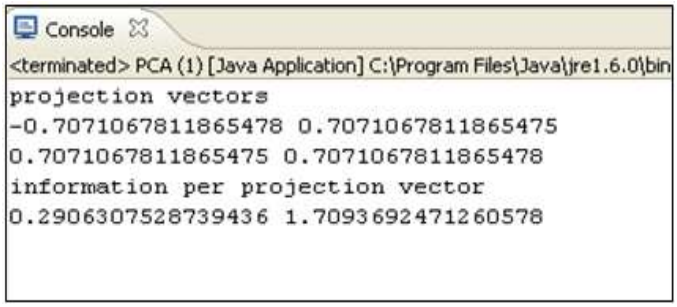

Fig. 7. PCA Results in Java

c. DR techniques statistics in Java Netbeans 8.0

The preprocessed matrix obtained in Java is given as an input to SVD as well as PCA code in Java, the statistics observed can be seen in Fig.6.It only represents the summary after computation of both techniques.

Fig. 8. SVD vs. PCA in JAVA

 ISSN: 2321-8169
5752 - 5757

2) MATLAB computations (SVD and PCA)

a. SVD Matrix with Computation time in MATLAB

MATLAB inbuilt function **[U,S,V] = svd(X,0)** is usd, where X indicates the input preprocessed matrix, whereas 0 signifies economy size decomposition. A dimensionally reduced dataset was produced in computational time 1.8513 seconds (Fig. 9.).

Fig. 9. SVD result in MATLAB

d. PCA Matrix with Computation time in MATLAB

DR toolbox was used to get access to the inbuilt functionality of MATLAB for PCA, the function provided by DR toolbox is **[mapped_data , mapping] = compute_mapping (data , method, # of dimensions, parameters).** A dimensionally reduced dataset was produced i.e. mapped data in computational time 1.8513 seconds (Fig. 10.).

Fig. 10 PCA result in MATLAB

3) Cluster formation in Java and MATLAB

a. Java Clusters (K-means algorithm)

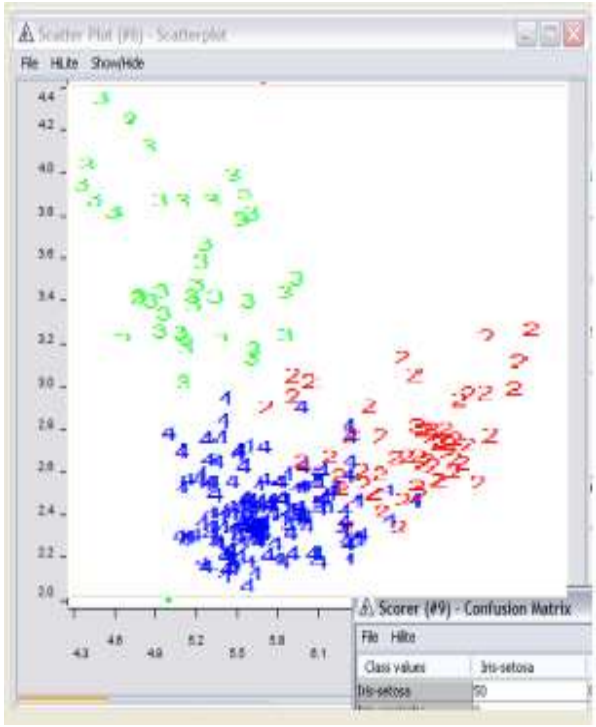

Fig. 11. K-means Cluster formation in Netbeans 8.0 (Java)

b. MATLAB Clusters (K-means algorithm)

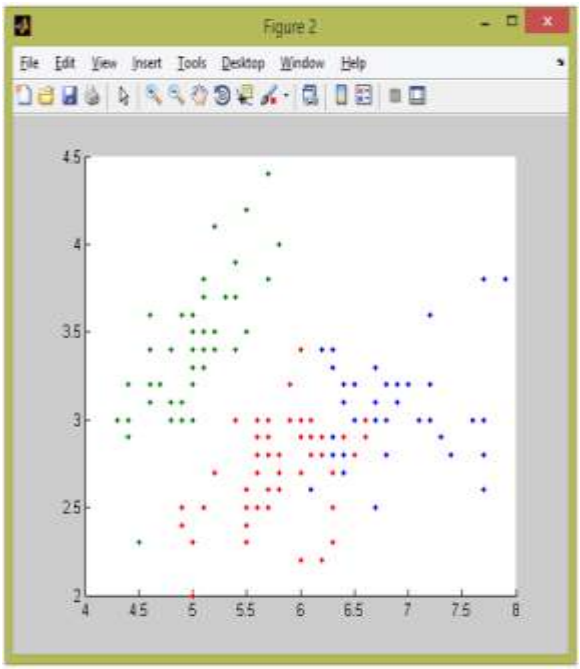

Fig. 12. K-means Cluster formation in Netbeans 8.0 (Java)

V. EXPERIMENTAL ANALYSIS

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 . Also, if the approaches for implementation of system i.e programming based and the tool based are compared then it is significantly noticed that the outcomes yielded by tool based approach i.e through MATLAB proved out to be better than that of programming based approach i.e through Java in Netbeans8.0.

TABLE I
COMPARISON OF Different Approaches For Expert IR

Approaches	SVD time (sec)	PCA time (sec)	Clustering
Tool Based (MATLAB)	1.8513	1.1524	Well separated and accurate
Programming Based (Java in Netbeans8.0)	3.157	1.780	Scattered and not accurate

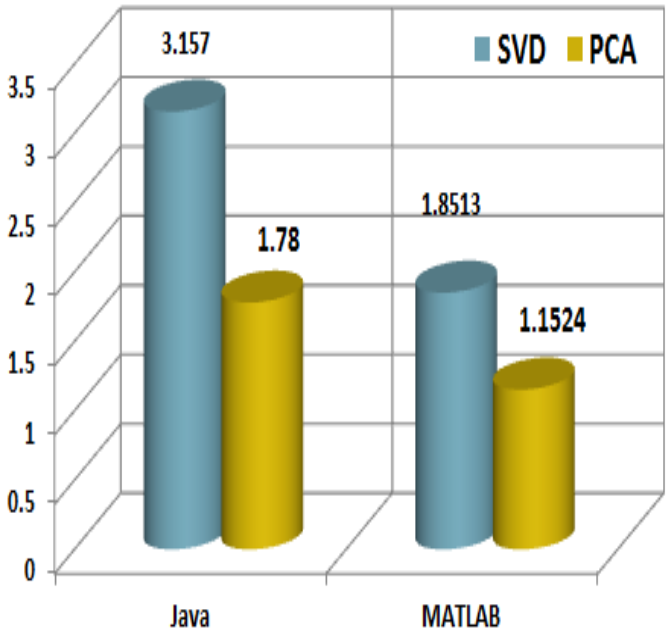

Fig. 13. DR techniques in Java vs. MATLAB upon computational time

VI. 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 information retrieval through clustering. Users can make comprehensive choices among various available dimensionality reduction techniques referencing this study. The main objective is to achieve best performance of 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. This research has shown comparable results of available techniques on C50test dataset and a better approach for relevant Information Retrieval system. Also a detailed analysis of various approaches like tool based approach using MATLAB and programming approach using Java is represented. Thus, showing the same system in different ways, wherein the programming approach allows us to develop a desired system in our feasibility and independence to program the system with modifications and in different ways ; on the other hand the tool based approach provides inbuilt functionality to implement various modules in the system with less control of the user as it works in accordance with its predefined functions. Thus, tool based system can be thought of as an easy implementation to obtain better results. The future scope for this research would be development of a user interface for relevant document retrieval, hence forming an expert IR system on high dimensional datasets.

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] NguyenHungSon,"Data Cleaning and Data Preprocessing".Lei Yu Binghamton University,
[3] Reduction for datamining Techniques, Applications and Trends", Jieping Ye, Huan Liu,Arizona State University.
[4] 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 Miningusing 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, E lizabeth 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] http://archive.ics.uci.edu/ml/datasets/Reuter_50_50

[14] https://sites.google.com/site/dataclusteringalgorithms/ k-means-clustering-algorithm