An Improved Document Clustering Approach Using Weighted K-Means Algorithm

[1] **Megha Mandloi; [2] Abhay Kothari**

[1] Computer Science, AITR, Indore, M.P. Pin 453771, India

[2] Computer Science, AITR, Indore, M.P. Pin 453771, India

Abstract - Now in these days digital documents are rapidly increasing due to a number of applications and their data generation. Such kind of data is very valuable for different application for business intelligence projects. But the key difficulty is the format of data and their domain identification. On the other hand not any kind of fixed pattern is available on such kind of data. Therefore supervised learning technique is not suitable for recognizing text pattern in such data sources. In order to solve these issues a new text clustering technique is introduced in this work. The proposed data model includes the following contributions (1) preparation of dataset, (2) designing two phase pre-processing technique (3) feature computation using weighted technique (4) feature selection and vectorization process (5) cluster analysis using k-means clustering and domain knowledge. Additionally, in order to provide justification and performance improvement over traditional tf-idf and k-means based text clustering technique is used. A number of experiments are performed, with randomly selected patterns (text documents with different group identifiers). In this context comparative performance results are prepared with help of most suitable values. According to comparative performance study 4-5 % clustering accuracy is improved. And the memory consumption is also preserved by 3-5%. But the implementation leaked with higher time consumption (approximate 10-30 MS (mean performance)) as compared to traditional clustering algorithm, due to additional feature extraction and selection process. Overall performance is adoptable as compared to traditional algorithm. Thus, it suggested applying such algorithm where higher accuracy is preferred for cluster analysis.

Keywords – *Text clustering, K-means, Dataset, Clustering, Data mining, Feature extraction*

1. Introduction

The text mining is a classical domain of research and application development. Now in these days number applications are usage the concept of text mining for business analytics and others. In this presented work the text mining techniques are studied and a new efficient and accurate technique is developed.

The data mining techniques are used for analyzing data, in order to find significant information among available raw data. These techniques are developed through the computational algorithms that help to automate processes required to analyze the data. According to the nature of data and their application requirements the different kinds of algorithms can be applied on the data.

For example if the data has some predefined classes and pattern samples then classification algorithms can be implementable, if there are some transactional sets are available and need to find frequent patterns then association rule mining is performed, or if the data has no class labels and need to find the different existing categories or groups on data then clustering is performed.

In this presented work the key aim is to study about clustering techniques. The clustering approaches are not much accurate because of their unsupervised nature of processes. Additionally the clustering approach can be applicable on text documents for finding their clusters more accurately. The clustering algorithm on text data is complex task, additionally achieving precise outcomes from the clustering over text data is also a complicated task. Therefore the key aim of the work is investigate about the different text clustering approach to enhance the traditional k-means clustering for text document clustering. In order to enhance the current clustering technique for text data the proposed work is intended to develop a improved weighted k-means clustering approach for precise outcome.

2. Objectives

The main aim of the proposed work is to find an accurate clustering scheme for text Therefore an improved k-means clustering technique for text clustering is proposed in this work. Additionally to solve the complexity of clustering, the following objectives are established for computation.

- **Study of text clustering technique:** in this phase the different clustering techniques which are frequently used in data mining tasks are studied. Additionally the most promising technique is recovered for further studies.
- **Study of different improvements on text mining approaches:** in this phase the different clustering improvement techniques are learned from the early literature review. Additionally adoptable technique for text clustering is obtained.
- **Design and implementation of the improved clustering technique**: in this phase a new clustering technique is designed using weighted technique for making more precise evaluation of text data. Additionally their implementation using suitable technique is performed.
- **Performance study of the proposed approach**: in this phase the proposed data model is evaluated for finding the improvements on existing clustering technique and their comparative outcomes are demonstrated.

The proposed work is based on cluster computing for text data. Therefore the key goal is established as design and development of improved clustering scheme, after successfully implementation of proposed task the following outcomes are expected.

- An improved approach of k-means clustering for making accurate document clustering using weighted technique
- A comparative performance study with k-means clustering and strength evaluation of the proposed methodology
- A new technique for document domain identification with less resource consumption (running time) as compared to traditional document clustering approach

3. Problem Definition and Proposed Solution

The rapidly increasing demand of digital data continuously increases the volume of digital documents. Processing and categorization of such significant amount of data for utilizing with different applications are a complicated task. In this presented chapter the digital document clustering technique is a main area of concern, therefore the key issues and their solution is introduced in this chapter.

3.1 System overview

Clustering is an unsupervised technique of learning and pattern recognition. Basically the learning is performed to the algorithm for teach them how to analyses less quantity of data to predict their actual pattern. As discussed the clustering is an unsupervised learning approach therefore the clustering of data not need to have any predefined classes. According to the data objects and their internal pattern similarity the algorithm decides the data object groups automatically. In this presented work the document clustering technique is investigated. Therefore a keen literature is collected where a number of techniques for cluster analysis are available. Among them the partition based clustering approach is a most popular technique for data analysis. Additionally in popularity the k-means clustering is a most frequently used algorithm in partition based clustering.

On the other hand the clustering techniques are also affected by the nature of data, for example if the dataset is available in structured format than the direct clustering approach is applicable to the datasets. Additionally if the data is not in structured format then strong pre-processing techniques and feature computation techniques are required to reduce the amount of efforts for performing the clustering. In this presented work the key aim is to provide the document clustering approach by enhancing the feature selection technique for text identification. In order to understand the core concept and proposed effort this chapter includes the detailed problem understanding and their solution for optimizing the performance of traditional document clustering approach.

4. Problem domain

According to observations and the evaluation of literature the following key issues and challenges are addressed for enhancing the traditional text clustering technique.

1. the length of the text documents are not similar therefore the evaluation of individual text contents needs a significant amount of computational resources
2. the feature extraction from the different documents are different in nature and length thus the similarity measurement of one data object to other object is a complex task
3. cluster formation of the documents need to select some centroids for accurate group formation, but random and fluctuating centroid selection in text documents can increase the process running time and their clustering accuracy
4. similarity approximation in text mining need to compare text document with their significant features but the directional information on similarity is computed yet for optimizing the performance of clustering

Solution methodology

In order to design an accurate and efficient clustering technique for text classification a new methodology for cluster analysis is proposed. The proposed technique's key components are listed using figure 3.1 additionally their sub-components are explained in detail. The process of the entire system is sub-divided into two major modules first the training and second the testing. But the proposed work is an unsupervised learning technique thus the training is not an appropriate term. Therefore the training process here termed as centroid selection process. Additionally the cluster formation process is termed here as the testing process.

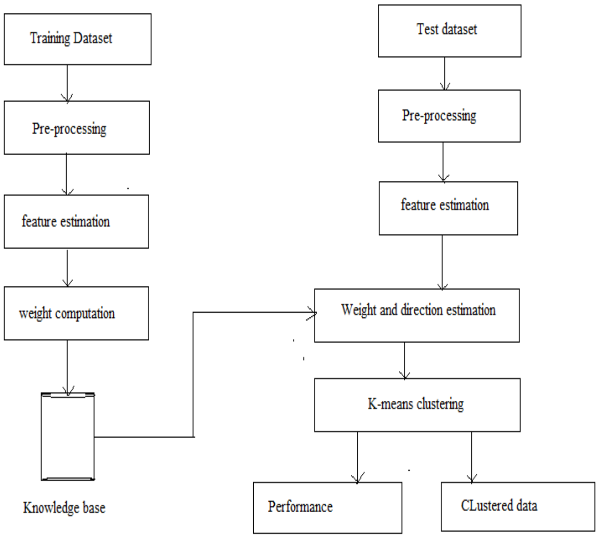

Fig 1: proposed document clustering system

Training dataset: in order to provide accurate data analysis the stable and accurate centroid selection process is required. Therefore the two different sets of data are used for training and testing purpose. In training process the data set is organized in terms of directories and their sub-directory manner. In figure the training dataset and their organization is represented. The root directory of the data is produced to system as the initial input for preparing the centroids. In addition of that to identify the patterns on the data the sub-directories is pre-labeled with the subjects or group names. In other words after or during clustering the test data can be recognizable in these specified sub-domains or group names. The group directories may contain one or more files for providing significance about the domain names. During process the key feature are need to approximate using these files for identify the group name or domain name.

Testing dataset: that is the secondary input to the system. That is also organized in form of directory but it contains a list of documents which are needed to be identifying into the groups which are learned by algorithm. Therefore a mix set of all groups file are prepared as the test set, and after algorithm's learning this dataset is produced to cluster the documents.

Pre-processing: the pre-processing is a process that identifies and removes the noisy content from the input datasets for learning. According to its name that process is applied before implementing the algorithm on the actual data. In the document mining the nature of pre-processing can be different from the other structured data mining techniques. In this context the two phase pre-processing technique is applied on the training data.

Removal of special characters: that is the first phase of pre-processing, in this phase the special characters from the entire text is removed (i.e. , " " / = +) % # @ and other similar). That process also helps to reduce the data for further data model building and the pattern recognition.

Stop word removal: in order to prepare sentences some of the words are frequently used such as (is, am, are, this, that) and others. Additionally these words are not having much significant for identification of any subject or group names therefore these data is also need to be removing from the text input.

Feature estimation: the volume of text data is always higher therefore the word to word comparison between a number of documents is a complex issue. Therefore to limit the amount of data for the comparison and other purpose the significant keywords are approximated from the text documents. Additionally these keywords are termed as the features of the text contents. In order to approximate the features from the available domains the two different features are estimated as:

Word frequency: this feature is computed for the individual word basis for the entire text domain or group name. For instance a group name "Data mining" contains 2 files which contain the total 1000 word by counting both the available files in this domain. Additionally need to compute the word frequency for word "Classification" which appeared in total of 10 times from both the documents. Then the word frequency is approximated by the following formula.

$$W_f = \frac{word\ occerence}{total\ words} \qquad (1)$$

Therefore the word classification's frequency is given by 10/1000.

Word importance: this feature helps to identify how important a word is for defining a group. Therefore that is computed on the basis of the word and the amount of sentences present in available documents. For example in the previous example, the data mining group contains total 100 sentences. And for constructing the sentence the classification word appeared in 10 different sentences then the word importance is computed on the basis of the following formula:

$$W_i = \frac{word\ found\ in\ sentances}{total\ sentances} \tag{2}$$

Therefore the classification word has the 10/100 scale of importance.

Weight computation: the weight computation helps to select the features from the total computed features from the total computed features. In addition of that, this phase also helps to regularize the length of estimated features. The trimmings of features are also termed as vectorization process. Therefore first the weights are computed for all the estimated tokens in the given domain of group according to the available files. The weigh computation is performed by using the following formula:

$$W = W_f * W_i \tag{3}$$

Now after computing the weights for all the computed tokens or words, the data is need to be select for ***vector*** development. This process is required because the length of the all the documents are not equal, additionally the computed number of features for all the documents are also not similar. Thus a common vector format is required to implement the clustering algorithm. In this presented work the length of vector is kept 50. In this context only those top 50 features are keep preserved which are having higher weights.

Knowledge base: that is the structured organization of the training documents. That is used to store the computed feature vectors into the database for utilizing the knowledge to identify any test document's subject or domain or group. Therefore the table shows the basis organization of knowledge.

Table 1: knowledgebase

Group Name	File name	Weighted tokens	Token weights

Weight and direction estimation: this process is taken place during the testing of the learned algorithm. Therefore first the test dataset is produced into the system which is evaluated according to the pre-processing phase and feature computation. After computing the features the entire features weights are combined using knowledge base information. Using this feature vector's direction and likelihood is approximated.

K-Means clustering: finally the traditional k-means algorithm is implemented for cluster the entire text documents available in test dataset. In this context the predefined centroids are produced as the group features are demonstrated. The classical k-means algorithm is given using table.

Table 2: K-means Algorithm

Input: N objects to be cluster (xj, xz … xn), the number of clusters k;
Output: k clusters and the sum of dissimilarity between each object and its nearest cluster center is the smallest;
Process: 1. Arbitrarily select k objects as initial cluster centers$(m_1, m_2, \dots, m_k)$; 2. Calculate the distance between each object Xi and each cluster center, then assign each object to the nearest cluster, formula for calculating distance as: $\mathrm{d}(\mathrm{x_i}, \mathrm{m_i}) = \sqrt{\sum_{\mathrm{j}=1}^{\mathrm{d}} (\mathrm{x_i} - \mathrm{m_{j1}})^2}, \mathrm{i} = 1 \dots \mathrm{N}, \mathrm{j} = 1 \dots \mathrm{k}$ $d(x_i, m_i)$is the distance between data i and cluster j. 3. Calculate the mean of objects in each cluster as the new cluster centers, $m_i = \frac{1}{N} \sum_{j-1}^{n_i} x_{ij}, i = 1,2, \dots, K$ N_i is the number of samples of current cluster i; 4. Repeat 2) 3) until the criterion function E converged, return $(m_1, m_2, \dots, m_k)$ Algorithm terminates.

Performance: in this phase on the basis of the clustered data performance of algorithm is computed. In this work for computing the performance accuracy, error rate, time and space complexity is measured.

Clustered data: the clustered data is termed for the obtained predictive outcomes for the input files as their group names which are need to be approximated. Therefore this phase results the group names for all the input test dataset.

5. Proposed Algorithm

In the previous section the entire system design and proposed system architecture is demonstrated. In this given system model two basic and important modules are implemented for accurately recognizing the document patterns. Therefore in order to demonstrate the entire clustering process with their training and testing phase the two algorithms is included in this section. Table shows the training algorithm and table demonstrate the testing algorithm.

Table 3: Training model

```
Input: training Dataset D
Output: Knowledge base K
Process:
  1. $R_d[Gp] = readDataSet(D)$
  2. $for\ (i = 0; i \leq R_d.length; i++)$
       i. $for(j = 0; j \leq Gp.fileCount;\ \ j++)$
            $Wf[\ \ ] = ComputeFrequency(Gp.file(j))$
            $Wi[\ \ ] = ComputeImportance(Gp.file(j))$
            $W[\ \ ] = ComputeWeight(Wf, Wi)$
            $V[] = computeVector(W[])$
       ii. $end\ for$
       iii. $K.append(V)$
  3. $end\ for$
  4. Return K
```

Table 4: Testing model

```
Input: knowledge base K, test dataset D
Output: clustered Data C
Process:
     $R_t = ReadTestData(D)$
  1. $for(i = 0; i \leq R_t.length; i++)$
       a. $C = kmean.Docluster(K,\ R_t^i,\ K.GroupNames)$
  2. End for
  3. Return C
```

6. Result Analysis

After successfully implementation of the proposed concept this chapter delivers the measured results. These results are computed during the different experiments and the most valued results are selected for demonstration. In addition of that a traditional k-means based clustering for text data is also compared with the td-idf based concept for performance justification.

Time consumption

The time consumption of the algorithm is also termed as time complexity of the algorithm. That is an amount of time which is required to complete the required clustering of documents. That can be computed using the time difference between algorithm initiation time and completion of the algorithm execution.

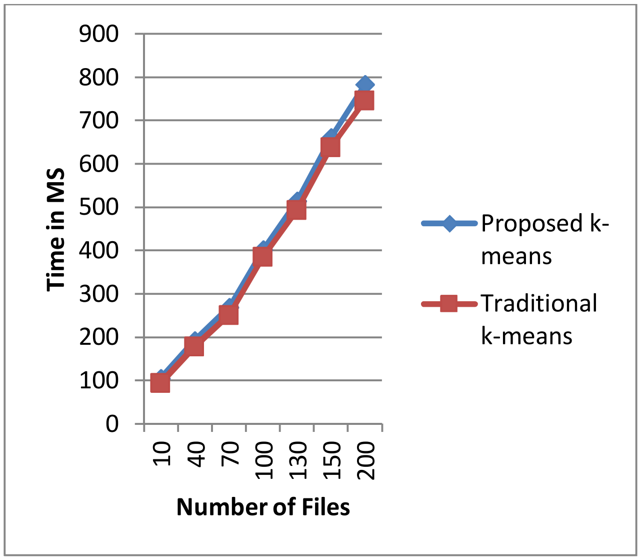

Fig 2: Time consumption

Table 5: Time consumption

S. No.	Data set size (number of files)	Proposed k-means	Traditional k-means
1	10	104	94
2	40	192	179
3	70	269	251
4	100	402	385
5	130	514	493
6	150	661	639
7	200	783	747

The required time for computing cluster is given using above figure and table. The figure contains the line graph for the algorithm computational time. Therefore in X axis the amount of files which are required to cluster is given and the corresponding amount of time is given in Y axis. The results demonstrate the required time is increases when the number of files for computation is increases. Thus the time consumption is depends on the number of files for clustering. Additionally the time is adoptable because after training the testing not requires huge amount of time for process the documents in the specific domain. As compared to the traditional technique of document clustering the proposed technique consumes higher time. Because the traditional technique computes only a single feature for approximating the clusters of the data, on the other hand the proposed technique need to features for clustering.

Memory consumption

Memory consumption is termed as space complexity of system. Thus it can be defined as the amount of main memory required to process the input training set. To compute the space complexity of any algorithm the difference between total assigned memory for a process and total free size of assigned memory is computed.

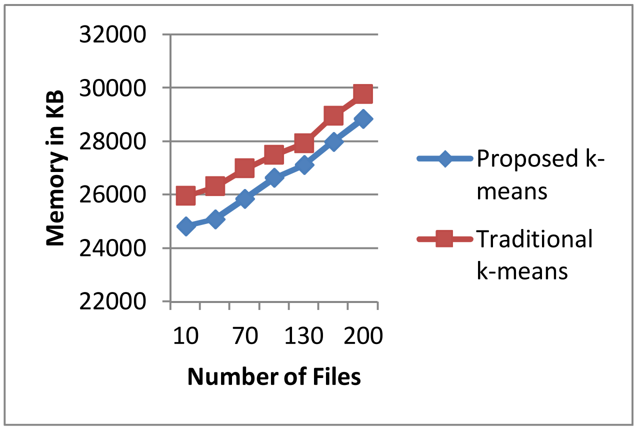

Fig 3: Memory usage

Table 6: Memory usage

S.No.	Parameters	Proposed (mean)	Traditional (mean)
1	10	24827	25933
2	40	25083	26295
3	70	25853	26971
4	100	26635	27485
5	130	27116	27918
6	150	27981	28942
7	200	28847	29751

The memory usage of the clustering algorithm with increasing amount of file size is demonstrated using above figure and table. The figure contains the graphical representation of the table data. To demonstrate the performance using the line graph the X axis contains the number of files for training and the Y axis contains the required main memory or computational space for the system. According to the experimental results the memory usage is also depends on the amount of files for processing. Thus as the number of files for processing is increases the required memory space is also increases in the similar ratio. In order to comparison with the traditional k-means clustering algorithm the proposed algorithm is memory resource preserving approach. Because the feature in the proposed technique is normalized with the help of regular size, on the other hand in traditional approach the irregular size of vectors are used for identifying the domain knowledge. Therefore the proposed technique consumes less memory during the pattern recognition.

Accuracy

The clustering an unsupervised learning technique therefore the accuracy is computed during the same time as the training performed. In this context the accuracy of the clustering algorithm is the amount of that is distinguished correctly during the clustering as the algorithm trains. Therefore the computation of accuracy is given using the following formula.

$$\text{accuracy }\% = \frac{\text{total correctly identified samples}}{\text{total input samples}} X100 \qquad (4)$$

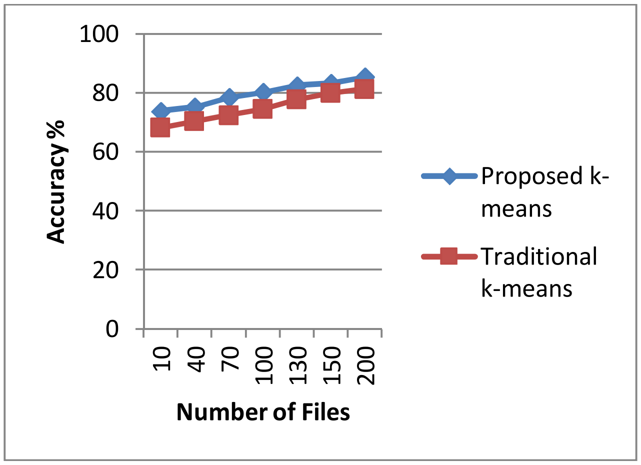

Fig 4: Accuracy

Table 7: Accuracy

S.No.	Parameters	Proposed (mean)	Traditional (mean)
1	10	73.85	68.32
2	40	75.29	70.47
3	70	78.52	72.53
4	100	80.19	74.61
5	130	82.57	77.84
6	150	83.24	80.05
7	200	85.48	81.39

The performance of the proposed algorithm in terms of accuracy is demonstrated using figure and table . The table shows the numerical values of performance in terms of percentage with increasing number of files. Additionally the graphical representation of the collected experimental values is given using above figure. To simulate the performance using line graphs the X axis contains the number of files produced as input to the system and the Y axis shows the accuracy obtained during experiments. According to the obtained performance the accuracy of the system needs a collection of significant knowledge therefore as the amount of knowledge in data base is increase the accuracy of the system is also increases. Because the correctly recognition of a text document in a domain needs to have the entire knowledge about the domain. The comparative performance of both the algorithms proposed and traditional k-means clustering for document clustering is given using figure and table. in this graph the red line shows the performance of traditional approach and proposed approach is given by blue line. According to the experimental results the proposed algorithm enhances the performance of the algorithm as compared to traditional technique. Thus the model for document clustering is much suitable for applications.

Error rate

The error rate of the algorithm demonstrates the amount of data which is not accurately identified during testing of algorithm. That is sometimes also termed as the misclassification rate in classification algorithms. To compute the error rate of algorithm the following formula can be used.

$$\text{error rate} = \frac{\text{misclassified samples}}{\text{total sample}} \text{X}100 \qquad (5)$$

Or

$$\text{error rate} = 100 - \text{accuracy} \qquad (6)$$

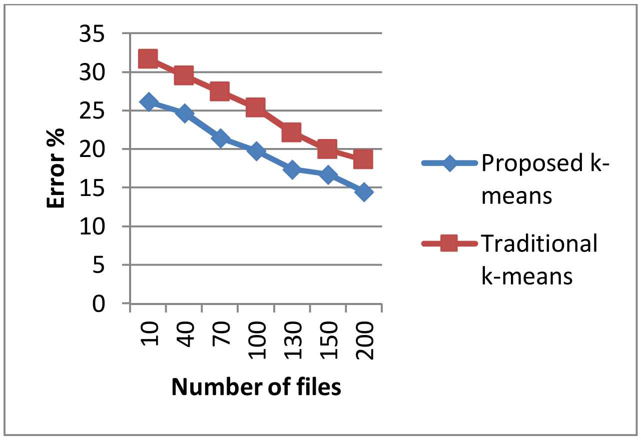

Fig 5: Error Rate

Table 8: Error Rate

S.No.	Parameters	Proposed (mean)	Traditional (mean)
1	10	26.15	31.68
2	40	24.71	29.53
3	70	21.48	27.47
4	100	19.81	25.39
5	130	17.43	22.16
6	150	16.76	19.95
7	200	14.52	18.61

The error rate of the proposed document clustering technique is demonstrated in the figure and table. The error rate of the algorithm is decreases as the amount of data for cluster analysis is increases in data base after successful training. The representation of the error rate using line graph is given here, in this line graph the X axis contains the amount of files produced for training and testing. Additionally the Y axis shows the obtained performance of the algorithm in terms of percentage. According to the experimental results the proposed methodology improves the performance of document clustering with increasing amount of training data. The comparative performance of proposed and traditional document clustering approach using k-means algorithm is given using above table and figure. In order to demonstrate the performance of proposed technique blue line is used and for demonstrating traditional approach the red line is used. According to the obtained results the proposed technique improves their performance much effectively as compared to traditional k-means clustering technique.

7. Conclusions

The data mining and their techniques are well known method for automated data analysis. According to the nature of data, the data mining algorithms are applied for different kinds of pattern recovery from raw data. In this presented work the clustering technique is main area of study. The clustering is an unsupervised manner of data analysis, where the data objects are evaluated on the basis of their internal similarity and the user defined groups of data is prepared. But the clustering approaches are not much accurate that is needed to be improving for accurate pattern identification. With this motivation the proposed work is focused on document based cluster analysis technique. In order to deploy the document based clustering the k-mean algorithm is one of the most popular approaches, therefore in this presented work an improved document clustering approach using the k-means algorithm is proposed and implemented. The proposed approach includes two phase of clustering first learning with the predefined patterns or groups, and in next phase utilizing the domain information for performing the cluster for incoming documents. During the training process the proposed system implemented the noise reduction technique using the stop word removal and the special character removal technique. In next process the feature extraction technique is used where the two technique are implemented first is implemented on the basis of word frequency in a specified domain and secondly the importance of a word in a given domain. After feature extraction the feature selection technique is used in this phase the vector is prepared for regular length based features evaluation and finally the k-means clustering with the predefined domain knowledge is implemented for computing more accurate clusters.

To implement the proposed methodology the JAVA technology is used. Additionally for the performance evaluation and their comparative study the traditional tf-idf based k-means clustering algorithm is used. To differentiate the performance of proposed and traditional approaches the accuracy, error rate and time and space complexity of both algorithms are computed and reported. The obtained experimental observation based performance of both the algorithms is given using table.

Table 9: performance summary

S.No.	Parameters	Proposed (mean)	Traditional (mean)	Improvement (Aprox)
1	Accuracy	79.88 %	75.03 %	4-5%
2	Error rate	20.12 %	24.94 %	4-5%
3	Memory usages	26620.28 KB	27613.57 KB	3-5%
4	Time Consumption	417.85 MS	398.28 MS	-10 - — 30 MS

he give performance summary as in table shows the effective ness of the proposed technique over traditional technique. According to the obtained mean performance of the algorithms the proposed model is efficient for accuracy, error rate and memory consumption. But the methodology increases the time consumption due to additional feature extraction and their refinement approaches.

Future work

The core objective of the proposed research work is achieved successfully and their implementation and performance evaluation is also performed. According to the obtained outcomes on unstructured data the methodology is effective and accurate for working large datasets too.

Therefore the following future extensions are proposed for work: Implementation of the proposed model over the email servers for spam filtering and their pattern recognition Implementation of the model with the big data analytic for trending topic evaluation where the multiclass sentiments are available Implementation of proposed model with academic research and document organization in huge digital library

References

[1] Anwiti Jain, Anand Rajavat, Rupali Bhartiya, "An Efficient Modified K-Means Algorithm To Cluster Large Data-set In Data Mining", International Journal of Advanced Research in Computer Science and Electronics Engineering Volume 1, Issue 3, May 2012

[2] Gurjit Kaur, Lolita Singh, "Data Mining: An Overview", IJCST Vol. 2, Issue 2, June 2011, ISSN: 2229-4333(Print) | ISSN: 0976-8491(Online)

[3] "An Introduction to Data Mining: Discovering hidden value inyourdatawarehouse", http://www.thearling.com/text/dmwhite/dmwhite.htm

[4] Manoj and Jatinder Singh, "Applications of Data Mining for Intrusion Detection", International Journal of Educational Planning & Administration. Volume 1, Number 1 (2011), pp. 37-42

[5] M. Rajalakshmi, M. Sakthi, "Max-Miner Algorithm Using Knowledge Discovery Process in Data Mining", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 3, Issue 11, November 2015

[6] SMRITI SRIVASTAVA & ANCHAL GARG, "DATA MINING FOR CREDIT CARD RISK ANALYSIS: A REVIEW", International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR), Vol. 3, Issue 2, Jun 2013, 193-200

[7] Dipti Verma and Rakesh Nashine, "Data Mining: Next Generation Challenges and Future Directions", International Journal of Modeling and Optimization, Vol. 2, No. 5, October 2012

[8] Hemalatha A.M, Ms. M. Subha, "A STUDY ON PLAGIARISM CHECKING WITH APPROPRIATE ALGORITHM IN DATAMINING", INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS, Vol.2 Issue.11, Pg.: 50-58 November 2014

[9] "UNIT I – INTRODUCTION: DATA MINING", http://www.kvimis.co.in/sites/kvimis.co.in/files/lectures_desk/DMBI_UNIT_1.PDF

[10] Tanu Verma, Renu, Deepti Gaur, "Tokenization and Filtering Process in RapidMiner", International Journal of Applied Information Systems (IJAIS) – Foundation of Computer Science FCS, New York, USA Volume 7– No. 2, April 2014

[11] Vishal Gupta, Gurpreet S. Lehal, "A Survey of Text Mining Techniques and Applications", JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 1, NO. 1, AUGUST 2009

[12] Rene Witte and Qiangqiang Li, Yonggang Zhang and Juergen Rilling, "Text Mining and Software Engineering: An Integrated Source Code and Document Analysis Approach", e IET Software Journal, Vol. 2, No. 1, 2008

[13] Nanasaheb Mahadev Halgare, Dharmaraj V. Biradar, "MPROVED ALGORITHM ON DYNAMIC CLUSTERING USING METAHEURISTICS IN ADVANCE DATA MINING", International Journal of Enterprise Computing and Business Systems ISSN (Online) : 2230-8849 Volume 6 Issue 1 January - June 2016