IJESRT
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

PROPOSED ALGORITHM FOR CONTENT BASED IMAGE RECOGNITION USING ENHANCED K-MEANS CLUSTERING ALGORITHM

Bhushit Chandra Nema*[1]& Khushboo Mandhaniya*[2]
*[1]Electronics & Communication, Ujjain Engineering College, Ujjain (M.P.), India
*[2]Electronics & Telecommunication, Institute of Engineering & Technology (DAVV), Indore (M.P.), India

DOI: 10.5281/zenodo.1002701

ABSTRACT

The content based image retrieval (CBIR) is the well-liked and heart favorite area of research in the field of digital image processing.The key goal of content based image retrieval (CBIR) is to excerpt the visual content of an image directly, like color, texture, or shape. There are several applications of the CBIR technique such as forensic laboratories, crime detection, image searching etc. For the purpose of feature extraction of well-matched images from the database, a universal CBIR system utilizes texture, color and shape based techniques. In this presented work, we have offered an efficient approach for the content based image retrieval, where images are decomposed using the wavelet transform, it means that the image features are converted in the matrix form and a color feature data set is prepared. In this paper, we are proposing the algorithm with the help of that we can improve the image retrieval.

KEYWORDS: Image Retrieval, Clustering, Wavelet Transform, HaarWavelet Transform, Feature Extraction, K-Means Technique.

I. INTRODUCTION

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 applicableto text documents for finding their clusters more accurately. The clustering algorithm on text data is acomplex task, so, achieving precise outcomes from the clustering over text data is a complicated. Therefore the principal aim of the work is to investigate about the different text clustering approach to enhance the traditional k-means clustering for text document clustering. Toenhance the current clustering technique for text data the proposed work is intended to develop an improved weighted k-means clustering approach to get theprecise outcome.

The main aim of the proposed work is to find an accurate clustering scheme for text clustering. 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.

a. 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 techniqueis recovered for further studies.
b. Study of different improvements on text mining approaches: in this phase, the different clustering improvement techniques are learned from the early literature review. Additionally, anadoptabletechnique for text clustering is derived.
c. Design and implementation of the improved clustering technique: in this phase, a new clustering technique is designed using theweightedtechniqueto make amore precise evaluation of text data. Additionally, their implementation using suitable techniqueis performed.
d. Performance study of the proposed approach: in this phase, the proposed data model is evaluated to find the improvements on existing clustering technique and their comparative outcomes are demonstrated.

II. PROBLEM DOMAIN

Clustering is an unsupervised techniqueof learning and pattern recognition. Basically the learning is performed on the algorithm toteach them how to analyze less quantity of data to predict their actual pattern. As discussed the clustering is an unsupervised learning approach, therefore, the clustering of data does 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 oftechniques for cluster analysis are available. Among them, the partition based clustering approach is a most popular technique for data analysis. Also, the k-means clustering is the most frequently used algorithm in partition based clustering.

According to observations and the evaluation of literature, the following key issues and challenges are addressedto enhance the traditional text clustering technique.

a. The length of the text documents are not similar therefore the evaluation of individual text contents needs a significant amount of computational resources
b. 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
c. 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
d. Similarity approximation in text mining need to compare text document with their significant features, but the directional information on similarity is computed for optimizing the performance of clustering.

III. PROPOSED 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 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.

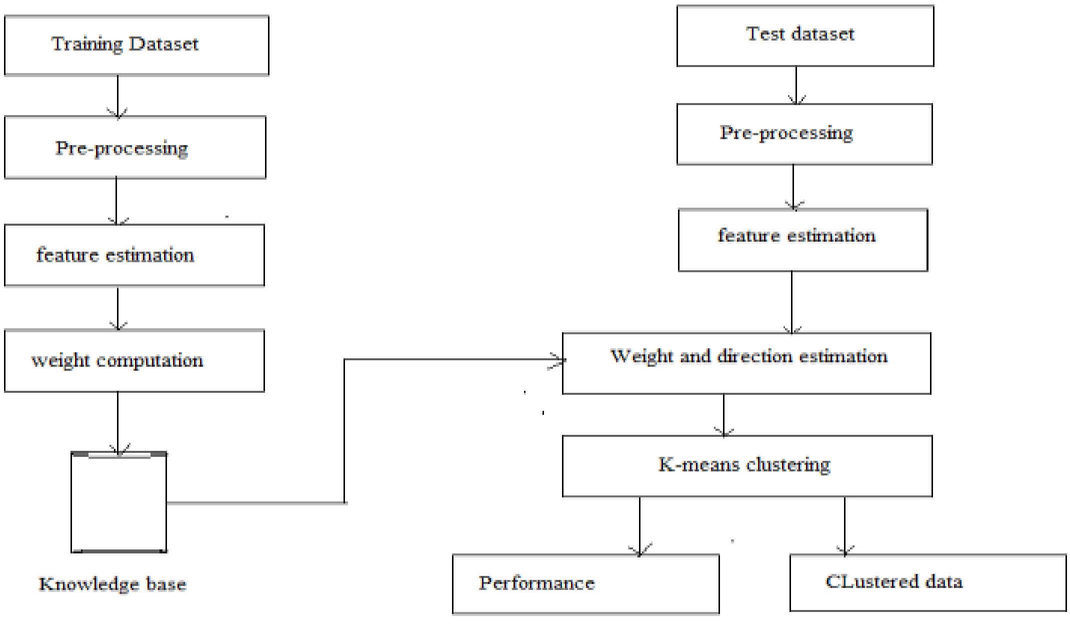

Figure1: proposed document clustering system

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 2 the training dataset and their organization are represented. The root directory of the data is produced to system as the initial input for preparing the centroids. Furthermore, to identify the patterns on the data, the sub-directories are 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 needs to approximate using these files for identifying the group name or domain name.

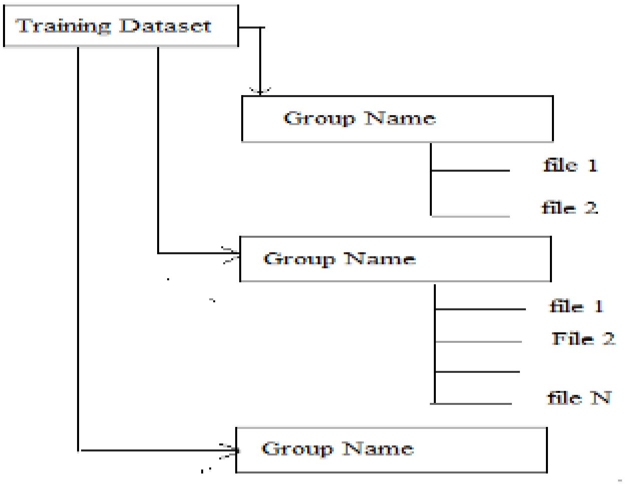

Figure 2: training set organization

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.

1. **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.
2. **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:

1. **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.

$$W_f = \frac{word\ occurence}{totalwords}$$

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

2. **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:

$$W_i = \frac{word\ found\ in\ sentences}{total\ sentences}$$

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:

$$W = W_f * W_i$$

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 3.1 shows the basic organization of knowledge.

Table 1 knowledge base structure

Group Name	File name	Weighted tokens	Token weights
Networking	Nw1	Topology	0.64
Computer Graphics	Cg1	Pixel	0.51

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 2. [9]

Table 2 k-means clustering

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 performance of algorithmis computed on the basis of the clustered data. In this work ofcomputing 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 needed to be approximated. Therefore this phase results in the group names for all the input test dataset.

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 3 shows the training algorithm and table 4 demonstrate the testing algorithm.

Table 3 training model

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

IV. CONCLUSION

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.

V. 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 in your data warehouse", 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

[6] 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

[7] 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

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

[9] 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

[10] 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

[11] 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

[12] 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

[13] L.V. Bijuraj, "Clustering and its Applications", Proceedings of National Conference on New Horizons in IT - NCNHIT 2013

[14] Patil Pravin Ishwar, Prof. Gajendra Singh, "A Novel & More Efficient Clustering Technique for Document Clustering Methodology", International Journal of Scientific Development and Research (IJSDR) Volume 1, Issue 4 [16] Yogita Rani and Harish Rohil, "A Study of Hierarchical Clustering Algorithm", International Journal of Information and Computation Technology ISSN 0974-2239 Volume 3, Number 10 (2013), pp. 1115-1122

CITE AN ARTICLE

Neema, B. C., & Mandaniya, K. (2017). PROPOSED ALGORITHM FOR CONTENT BASED IMAGE RECOGNITION USING ENHANCED K-MEANS CLUSTERING ALGORITHM. *INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY, 6*(10), 184-190.