International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 09 | Sep-2016 www.irjet.net p-ISSN: 2395-0072

Improved k-means Clustering for Document Categorization

Amandeep Kaur[1], Tarun Kumar[2]

[1]*M. Tech Scholar, Dept. of C.S.E, GIMT Kanipla, Kurukshetra University,Kurukshetra, India*
[2]*Assistant Professor, Dept. of C.S.E, GIMT Kanipla, Kurukshetra University, Kurukshetra, India*

Abstract - *Document categorization is used for sort the useful document and classifies the document by content. Document categorization is document classification. It is an approach of machine learning in the form of Natural Language Processing (NLP). Our goal is to assign one or more classes or categories to a document, which makes it easier to sort and manage. In our research dataset is used and read the documents. The special symbols, stemming, and stop words are removed. Lowercase conversion performed to reduce the time. The occurrence of repeated words also measured. The tf-idf also calculated for vector space model. We also predict the centers and finding out the nearest neighbor. For the evaluation of performance precision, recall and f-measure also calculated.*

Key Words: Document categorization, document clustering, k-means, ikmeans, data mining, clustering.

1. INTRODUCTION

Document categorization involves the use of descriptors and descriptor extraction. Descriptors are sets of words that describe the contents within the cluster. Document clustering is generally considered to be a centralized process. Examples of document clustering include web document clustering for search users.The application of document clustering can be categorized to two types, online and offline. Online applications are usually constrained by efficiency problems when compared to offline applications. The objective of document classification is to reduce the detail and diversity of data and the resulting information overload by grouping similar documents together [3].

Document classification or document categorization is a problem in library science, information science and computer science. The task is to assign a document to one or more classes or categories. This may be done "manually" (or "intellectually") or algorithmically. The intellectual classification of documents has mostly been the province of library science, while the algorithmic classification of documents is mainly in information science and computer science [2]. The problems are overlapping, however, and there is therefore interdisciplinary research on document classification. The documents to be classified may be texts, images, music, etc. Each kind of document possesses its special classification problems. When not otherwise specified, text classification is implied.

Documents may be classified according to their subjects or according to other attributes (such as document type, author, printing year etc.). In the rest of this article only subject classification is considered. There are two main philosophies of subject classification of documents: the content-based approach and the request-based approach.Document categorization is a problem in which task is to assign a document to one or more classes or classification [6].

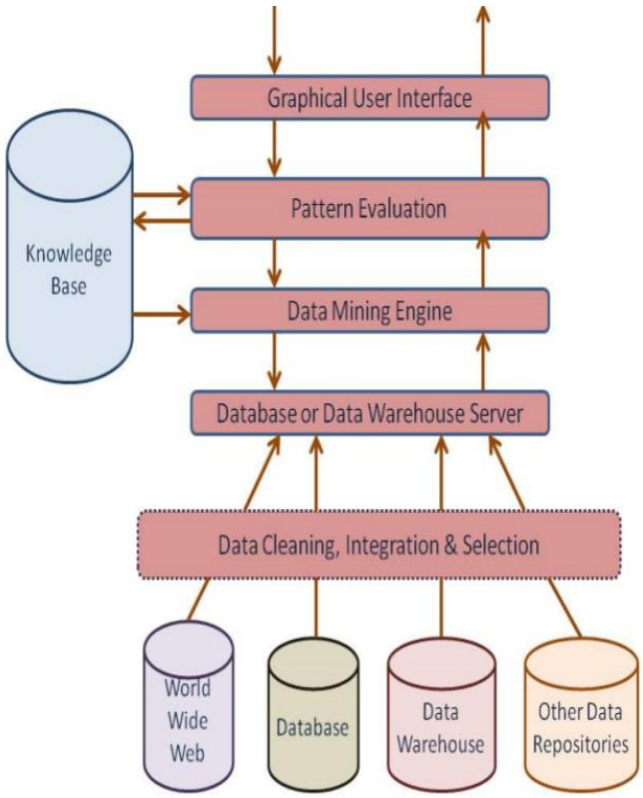

Figure 1- Data Mining Architecture

2. RELATED WORK

Victor Mireles, Tim O.F. Conrad [7] stated that a set of n binary data points, a widely used technique is to group its features into k clusters. We have presented analytical results

and a method that aid in the estimation of upper bounds to the sparsity of an overcomplete feature clustering. We believe the analysis presented here can guide us in providing more parsimonious interpretations of data. In this case where n < k, the question is how overlapping are the clusters becomes of interest. In this paper we approach the question through matrix decomposition, and relate the degree of overlap with the sparsity of one of the resulting matrices. We present analytical results regarding bounds on this sparsity, and a heuristic to estimate the minimum amount of overlap that an exact grouping of features into k clusters must have.

AmolBhagat, NileshKshirsagar, PritiKhodke, KiranDongre, Sadique Ali [1] stated that extracting useful information from large sets of data is the main task of data mining. Data streams are sequences of data elements continuously generated at high rate from various sources. Data streams are everywhere and are generated by the applications like cell-phones, cars, security sensors, televisions and so on. Partitioning data streams into sets of meaningful subclasses is required for proper and efficient mining of intended data.This paper addresses the issue of identifying the number of clusters by proposed penalty parameter selection approach. The performance parameters used are precision, recall, purity, G-precision, G-Recall.

HajarRehioui, AbdellahIdrissi, ManarAbourezq, FaouziaZegrari [4] stated that every day, a large volume of data is generated by multiple sources, social networks, mobile devices, etc. This variety of data sources produce an heterogeneous data, which are engendered in high frequency. Finding a compromise between performance and speed response time present a major challenge to classify this monstrous data. For this purpose, we propose an efficient algorithm which is an improved version of DENCLUE, called DENCLUE-IM. The idea behind is to speed calculation by avoiding the crucial step in DENCLUE which is the Hill Climbing step. Experimental results using large datasets proves the efficiency of our proposed algorithm.

Samantha Susan Mathew, Hafsath C A [5] stated that Cloud computing involves storage as well as usage of computer resources and services online. It has emerged as a promising area for data outsourcing. Since the users data are stored in servers controlled by cloud service provider there are always concerns about the security of the stored data. The approach also makes sure that the processing on the user side is minimized while updating the existing document set. The focus of the approach is to perform the bulk of the processing at the cloud server side which has high computation power. The confidentiality is ensured by using homomorphic encryption technique.

3. PROPOSED WORK

3.1 Dataset Reader:

Dataset reading performed by extractor and read the number of documents, the number of topics and the documents which are related to the specific topics. Reader read input text document and divide the text document into a list of features which are also called (tokens, words, terms or attributes).

A. Preprocessor:

Pre-processor processed the document words by removing special symbol, stop words, lower case conversion and stemming. Special symbols are the words which have their special meaning like <, >, _ etc. are removed. Stop words are like “an”, “the”, “a”, “and” etc. These stop words are removed from documents. Lower case conversion also performed to convert the all words of the documents in lower case. Stemming is the process of removing affixes (prefixes and suffixes) from features. For example: reads and reading are two words. After removing stemming “s” and “ing” are removed from the word read.

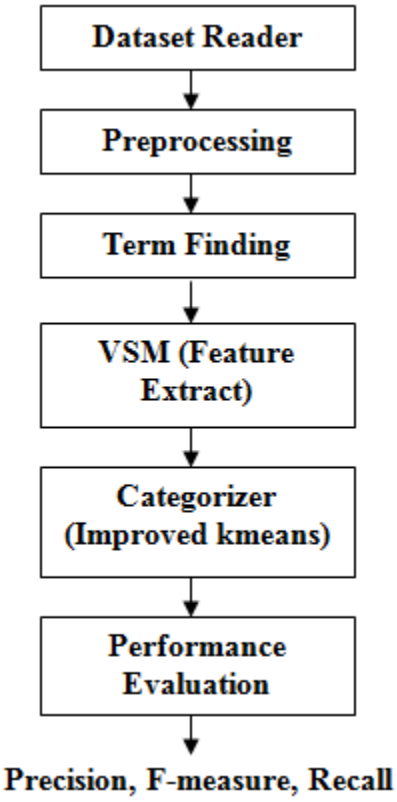

Figure 2- workflow of proposed work

3.2 Term Finding:

In term finding, we find the top fifty words in the documents. The words that occur again and again in the document are finding. First of all, we read the file to find out the occurrence of words. Note down the occurrence of repeated words. The repeated words are found out so the time complexity decreases.

3.3 Vector Space Model:

In vector space model each input text document is represented as a vector and each dimension of this space represents a single feature of that vector and on the basis of frequency of occurrence, weight is assigned to each feature in text document. This representation is called vector space

model. In this step, each feature is assigned to an initial weight equal to 1. This weight may increase based on the frequency of each feature in the input text document. Vector space model use feature extraction method which detects and filter only relevant features which are far smaller than actual number of attributes And this process enhances the speed of supervised learning algorithms.

TF-TDF term is used in vector space model for assigning weight to each feature. It determines the relative frequency of words in a specific document. For calculation, TF-IDF method uses two elements:
TF - term frequency of term in document (the number of times a term appears in the document)
IDF- inverse document frequency of term i (the number of documents where the term appears)

$$tf(t,d) = 0.5 + \frac{0.5 * f(t,d)}{\max\{f(w,d) : w \in d\}}$$

$$idf(t,D) = log\frac{N}{|\{d \in D : t \in d\}|}$$

$$tfidf(i,d,D) = tf(t,d) * idf(t,D)$$

3.4 Categorizer:

The categorizer is used to categorize the document. It includes the two parts i.e. center prediction and nearest neighbour.

1. Center Prediction: Before calculating centre prediction, a vocabulary is formed from the documents. Like, for one category a vocabulary of 30 frequent words are chosen same for category two and same for category three. These 90 words are quite frequent from each category and centre calculation is properly based on the frequent coming words and vocabulary. First 30 words are centre for one category next 30 words are centre for second category and next 30 words are centre for third category. Training of data is based on the centre prediction. Centres are predicted from intelligent vocabulary which contains top rated terms which reduce dimension and complexity and testing become easier.
2. Nearest Neighbour: In nearest neighbour we define our proposed algorithm i.e. ikmeans (improved kmeans). The ikmeans algorithm improves time complexity. The algorithm defines below:

Proposed Algorithm of improved k-means (ikmeans):
Input: VSM (data), k
Where k is the total number of clusters
Output: center, cluster
Steps:
1. Center prediction
   (i) Sort the data into ascending/descending order.
   (ii) Divide into k numbers.
   (iii) Calculate mean of columns for respective center.
2. Repeat
Assign each point *di* to the cluster which has the closest centroid;
Calculate the new mean for each cluster;
Untilconvergence criteria is met.

3.5 Performance Evaluation:

The performance evaluation uses the three methods to evaluate the result. The methods are f-measure, precision and recall. The formulas for precision, recall and f-measure are given below:

1. Precision:

$$precision = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{retrieved\ documents\}|}$$

2. Recall:

$$recall = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{relevant\ documents\}|}$$

3. F-measure:

$$F1 = 2.\frac{precision.recall}{precision + recall}$$

4. RESULTS

The proposed work is implemented on Intel inside® Core(TM) i3 Processor 2.20 GHz, Windows 10 and used matlab 7.10.0 platform and compare results with existing work. In this, the mini newsgroup dataset are used for performing the proposed experiment and achieving 100% result as compare to existing work.

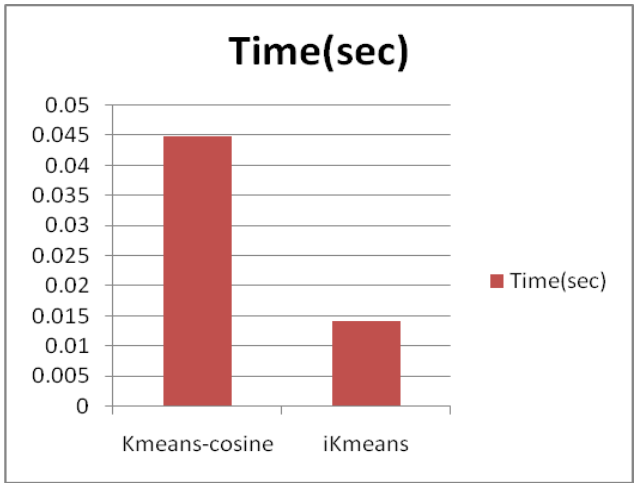

Fig 3- Time complexity comparison between kmeans and ikmeans

The figure 3 shows the time complexity of kmeans and ikmeans. Our proposed ikmeans algorithm reduced the time complexity 3.06% than kmeans algorithm.

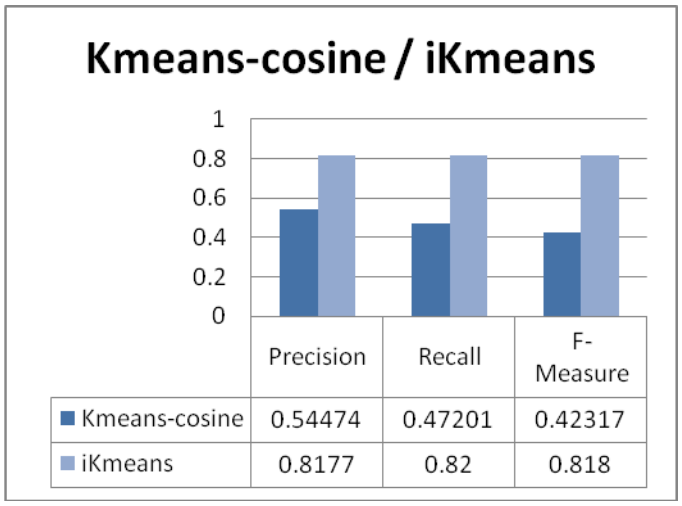

Fig 4- Existing and proposed algorithm comparison

The figure 4 shows the normalized value of precision, recall and f-measure. The precision of ikmeans is 27.296% greater than kmeans. The recall of ikmeans is 34.799% greater than kmeans. The f-measure of ikmeans is 39.483% greater than kmeans.

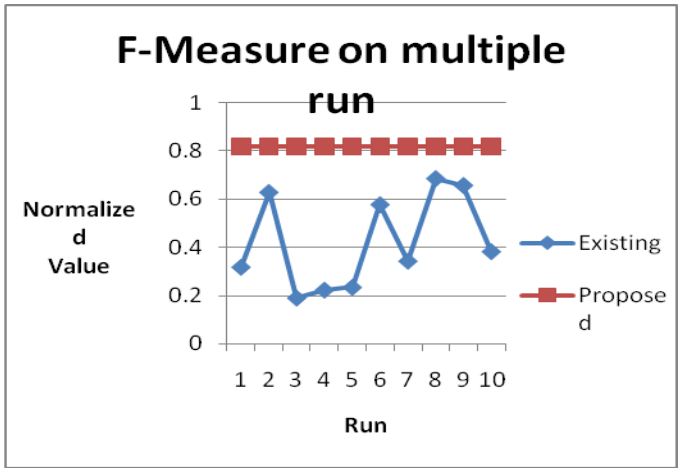

Fig 5-F-measure on multiple run

The figure 5 shows the comparison between existing and proposed f-measure on multiple run.

5. CONCLUSIONS

Document categorization is used to divide the documents into clusters. In a document same type of data is used. Our goal is to reduce the time complexity and increases the accuracy. In this paper, we performed document categorization on mini newsgroup dataset. Document categorization is document classification. It is an approach of machine learning in the form ofNatural Language Processing (NLP). We assign one or more classes or categories to a document, which makes it easier to sort and manage. We proposed a new algorithm named as improved kmeans (ikmeans). The ikmeans gives us better result than existing kmeans algorithm. The precision, recall and f-measure of improved kmeans achieve 27.296%, 34.799%, 39.483% better result than existing kmeans algorithm. In existing algorithm the cluster's center are chooses randomly. But in our proposed work, the center of cluster are fixed and calculated by using average value of mean. The result shows high accuracy in less time.In future the proposed work can be tested on multiple datasets as well as used with classifier.

REFERENCES

[1] AmolBhagat, NileshKshirsagar, PritiKhodke, KiranDongre, Sadique Ali "Penalty parameter selection for hierarchical data stream clustering", 2016

[2]https://en.wikipedia.org/wiki/Document_classification

[3] HeideBrucer, Gerhard Knolmayer, Marc-Andre Mittermayer "Document Classification Methods for Organizing Explicit Knowledge" 2001

[4] Hajar REHIOUI, Abdellah IDRISSI, Manar ABOUREZQ, Faouzia ZEGRARI "DENCLUE-IM: A New Approach for Big Data Clustering", 2016

[5] Samantha Susan Mathew, Hafsath C A "Aiding Effective Encrypted Document Manipulation Incorporated with Document Categorization Technique in Cloud", 2015

[6] Shaifali Gupta, Reena Rani "A Review of Document Categorization Technique", 2016
[7] Victor Mireles, Tim O.F. Conrad "Minimum-overlap clusterings and the sparsity of overcomplete decompositions of binary matrices", 2015

BIOGRAPHIES

Amandeep Kaur is M.tech. Scholar of Computer Science & Engg. Branch. Her area of interest is Data Mining.

Er. Tarun Kumar currently working as Astt. Prof. in CSE Department (GIMT, Kurukshetra) having experience of almost 210 yrs. In teaching. He has published more than 10 papers in national and international Journals and conferences. He has guided more than six M.tech. scholers. His area of interest is Data Mining, Wireless Sensor Networks.