Attribute Weighted K-means For Document Clustering

Monika Gupta[1], Kanwal Garg[2]

[1]M.Tech.(CSE) Scholar, Department of Computer Science & Applications, Kurukshetra University,
Kurukshetra-136119, Haryana, India
[2]Assistant Professor, Department of Computer Science & Applications, Kurukshetra University,
Kurukshetra-136119, Haryana, India

Abstract- *Document clustering has been one of the fastest growing research field for the past few decades. It has become an important task in text mining because of the tremendous increase in documents on the internet. All the organizations need the proper management of textual data. Document clustering is the unsupervised technique that helps to organize the similar documents into classes to improve retrieval. The paper explains the phases of document clustering and the improvement in document clustering using attribute weighted k-means to cluster the documents and to put the similar documents in the proper cluster. Experimental results shows that accuracy of proposed method is high compare to the basic k-means in terms of F-Measure and time complexity.*

Keywords: *Document Clustering, Cosine Similarity, F-measure*

1. Introduction

Clustering is the unsupervised classification in which there is no predefined class. Clustering technique helps to group the data objects into classes or clusters. A cluster is the collection of data objects such that the objects are similar to one another within the same cluster and dissimilar to the objects in different clusters. A good clustering method is one that produces high quality clusters with high intraclass similarity and low interclass similarity. Clustering is very useful in several machine learning and data mining tasks that includes information retrieval, pattern recognition, pattern classification etc[11]. The text document clustering is the act of grouping similar documents into clusters to better discriminate the documents belonging to different categories[6]. The main aim of clustering is to discover a natural grouping between a set of patterns, objects without the knowledge of class labels[7]. The clustering technique is very helpful in text domain where the objects that are to be clustered are of different granularities such as documents, paragraphs, sentences or terms. It plays an important role for organizing documents to improve retrieval and support browsing[3]. Document clustering is the automatic organization of similar documents into group's text extraction in an unsupervised manner for the fast information retrieval. The objective of clustering is to minimize the intracluster distance & maximize the intercluster distance[14]. The two major similarity criterions are distance based & concept based. Two or more objects belong to the same cluster if they are "close" according to a given distance. This is called distance based clustering. In conceptual clustering – two or more objects belongs to the same cluster if one defines a concept is common to all that objects[12].

The rest of the paper is organized as follows: The related work is discussed in section 2. Section 3 describes the proposed work. Section 4 describes the evaluation measures for the clustering. Section 5 describes the result analysis of proposed work. Finally, in section 6, the paper is concluded.

2. Related work

The term data clustering was first introduced in the title of a 1954 article dealing with anthropological data[9]. Abhilash C B et al. (2013) [] performed a comparative study for the clustering of data with the help of improved K-means algorithm. The standard K-means and the proposed K-means were compared and it was found that the improved K-means helps to produce optimal cluster and require less number of iterations. The clustering for both the algorithms was performed on the yeast and iris dataset. The improved K-means use the minimum spanning tree concept. For all the input data points, the undirected graph was produced and then from all the paths, the shortest distance was found out. Khaled M. Hammouda et al. [5] explains the two key parts of document clustering. The first one is phrase based document index model, the Document Index Graph that helps for the incremental construction of phrase based index of document set. The second one is incremental document clustering algorithm that forms its basis on maximizing the tightness of clusters by determining the pair wise documents similarity distribution inside cluster. Wael M. S. Yafoz et al. [17] focuses on the fact to group the textual documents based on the similarities. It also shows the importance of dynamic clustering for mining the frequent terms with included named entity.

3. Proposed work

Data set*:* The mininewsgroup20 is a very standard and popular dataset used for evaluation of many text applications, data mining methods, machine learning methods, etc[19]. The another datasets that are used are minidataset[20] and 20_newsgroups[19]. After dataset reading, the preprocessing of the documents is carried out to remove the irrelevant terms.

Tokenization & Preprocessing: In tokenization process, the document is split into stream of words by removing all the punctuation marks and by replacing tabs and other non-text characters by single white spaces. The preprocessing helps to remove stopwords, special symbols and the words that are irrelevant. The idea of stopword removal is to remove the words that carry little or no content information like articles, conjunctions, prepositions etc. Stemming helps to build the basic form of words like "removing" and "removed" can be used as "remove"[18]. All the irrelevant words or the words carry little information are discarded and the terms are reduced to their stem in this step. A total of 12304 terms from 53869 are extracted from mininewsgroup20 dataset and 14226 terms from total of 38462 is extracted from mini_dataset and 76178 terms from a total of 298912 terms are extracted from 20_newsgroup dataset after preprocessing.

Term Finder: After preprocessing, the term finder finds the terms from all the documents that have number of occurrence equal to or greater than threshold. The exclusive words from both the document categories are picked.

Input:

(i) Total number of terms
(ii) Threshold

Output: Term set

After this step the terms reduced to 826,798 and 2361 for mininewsgroup20, mini_dataset, 20_newsgroup respectively.

Feature Extraction: The set of features are produced by parsing each document. The process helps to remove noise and reduces the dimensionality of feature space. The widely used feature selection metric is term frequency and inverse document frequency.

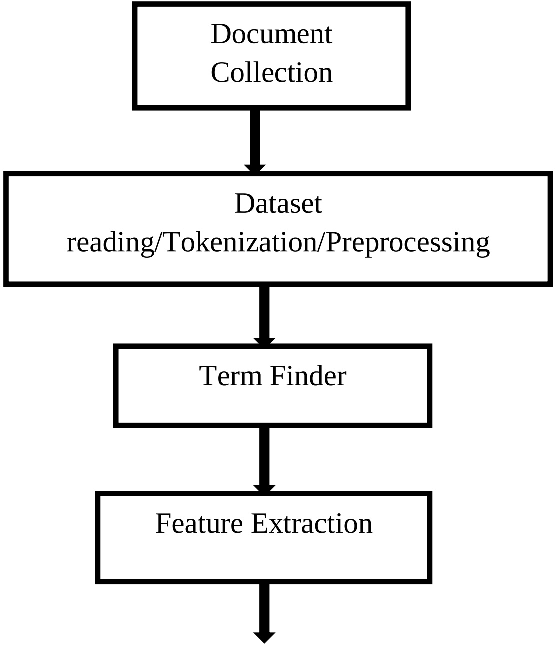
Term frequency finds out the weight of each word in each document. TF counts the occurrence of terms in a document. It is assumed that more frequent words are more significant. Weight for each term is calculated as: Wtt,d = tft,d* idf t

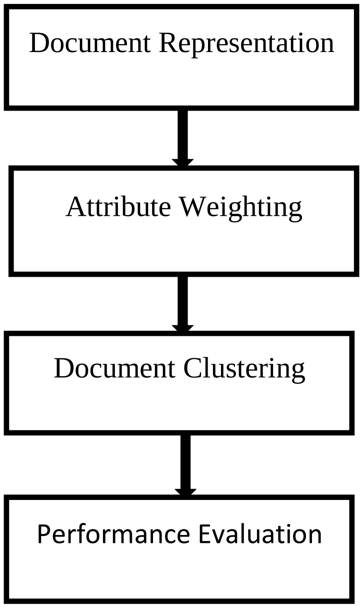
Inverse document frequency is: idf = log (N/ dft), where N represents the total number of documents, dft is the number of documents that the specific term appears[13].

Document Representation: After extracting the features, the document representation is done with the help of vector space model. The documents are represented as the vector of keywords. A collection of n documents with m unique words are represented by the m*n matrix[15].

	Term1	Term2	Term3	Termn
Doc1	2	3	3	2
Doc2	4	2	3	2

It may happen that VSM contains number of sparse entries. The terms that appear commonly in both the categories may produce difficulty. So, the sparse entries and the terms that appear equal number of times in both the categories are discarded. Then, the VSM acts as a input to the clustering phase.





Clustering: Document clustering treats VSM as input and produces the clusters as output. The proposed clustering method consists of two steps:

a) Attribute Weighting
b) Regular K-means

Attribute Weighting: The main task of attribute weighting method is to map the attribute according to their distribution in the dataset. The similar data within the each attribute is to gather altogether to increase the discrimination between classes[4].

Gainratio is about the probabilistic distribution of each word. By combining the weight based on gainratio with term frequency and inverse document frequency, we can assign a composite weight to each word in document[16].

If(gain(ai) ≥ threshold), the specific column is used for clustering.

The feature reduced the terms to 613,464 and 1413 for mininewsgroup20, mini_dataset and 20_newsgroups respectively.

The filtered vsm acts as input to the K-means on which regular K-means applied to form the cluster of documents.

The very first use of k- means algorithm is done by James McQueen in 1967. K-means is a partition method technique in which n objects are partitioned into k clusters, where k < n. When the data is clustered in k groups, the following condition will be fulfilled:

- Atleast one object in each group
- Each object belongs to exactly one group[8].

The distance measure that is used is cosine similarity. In cosine similarity, when documents are represented as term vectors, the similarity of two documents corresponds to the correlation between the vectors. This is quantified as the cosine of the angle between vectors[10]:

Given two documents $\vec{t_a}$ and $\vec{t_b}$, their cosine similarity is:

$$\mathrm{SIM}_{\mathrm{C}}(\vec{t_a}, \vec{t_b}) = \frac{\vec{t_a} \times \vec{t_b}}{|\vec{t_a}| \times |\vec{t_b}|}$$

K- means algorithm:

Input: D = {d1, d2, d3-----------, dn}//

Where D contains data objects

No. of clusters = k

Output: A set of k clusters

Step 1: Initialize: randomly select initial cluster centroid from data items.

Step 2: Assignment: assign each data point to the cluster having closest mediod based on cosine similarity measure

Step 3: Update: recalculate new mean for each cluster

Step 4: repeat step 2 and step 3 until the desired number of cluster is reached[8].

n data points are given as input to form clusters. K are the number of clusters to be formed. The algorithm randomly selects k centroids to form clusters, then the data points are assigned to the nearest cluster based on the cosine similarity measure. After then, the centroids are updated as center points that are obtained in first step. The process terminates when there is no cluster updation takes place[10].

Pseudocode:

Input: The input is the collection of text documents.

Output: k number of clusters.

Step 1: For every document, preprocessing will be carried out that helps in-

- Irrelevant stopword removal

- Special symbol removal
- Conversion of all the data in lowercase

Step 2: The term finder then finds the terms in each and every document having value equal to or more than threshold.

(i) Find frequent terms for all words

if tfi ≥ 3, add to set

(ii) For each termset, ts1, ts2, exclude the exclusive words

If tfi ≥ 2, then exclude

Step 3: Calculate tfidf and prepare a VSM that contains terms with their weights. The terms having equal number of occurrences in both the categories are discarded to avoid confusion.

Step 4: Find optimal attribute set using gain ratio and filter the vsm.

Step 5: Filtered vsm goes to the k- means for the clustering of documents.

Step 5: Finish

4. Evaluation measures

For the evaluation of quality of clustering, three quality measures are widely used for the purpose of document clustering. The first is F-measure that combines precision and recall.

Precision is the percentage of the documents that are correctly classified. **Recall** is the percentage of total documents that are correctly classified. The formula for Precision is TP/(TP+FP) and recall is TP/(TP+FN)
The precision and recall for cluster j with respect to a class i is defined as[5]:

P= Precision(i,j)= N_{ij}/N_j
R= Recall(i,j)= N_{ij}/N_i

Where N_i = Number of members of class i

N_j = Number of members of cluster j

N_{ij} = Number of members of class i in cluster j

F-measure for class i is given as:

F(i) = 2PR/**(**P+R)

The second measure is entropy. It provides a measure of "goodness" for unnested clusters. Entropy determines how homogeneous a cluster is. The homogeneity and entropy is inversely proportional i.e. higher the homogeneity lower the entropy. Cluster containing one object have 0 entropy[15]. The entropy of a cluster C_i with size n_i is defined as[2]:

$$E(C_i) = -\frac{1}{logc}\sum_{h=1}^{k}\frac{ni^h}{ni}\, log\left(\frac{ni^h}{ni}\right)$$

The third measure is purity that evaluates the coherence of a cluster, that is, the degree to which a cluster contains documents from a single category. Given a particular cluster C_i of size n_i ,the purity of C_i is formally defined as[1]:

$$\mathrm{P}(\mathrm{C_i}) = \frac{1}{n_i}\max_h(n_i^h)$$

In general, maximize the F-measure and purity and minimize the entropy in order to achieve high quality clustering.

5. Experimental Results

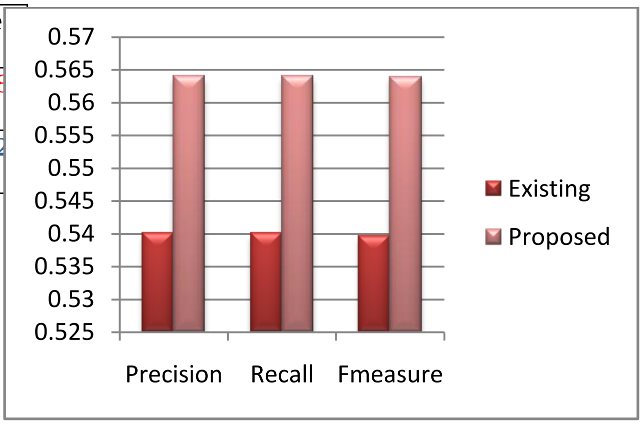
The work is implemented using NetBeans7.3, 4 GB RAM, in Windows 7, 64-bit operating system with i3 processor. The mininewsgroup20 and mini_dataset is used to evaluate the clustering performance. The mininewsgroup20 categories alt.atheism and comp.os.ms-windows.misc each having 100 documents, the mini_dataset categories positive and negative each having 25 documents and the 20_newsgroups categories rec.sport.hockey and rec.sport.baseball each having 500 documents is used to perform clustering.

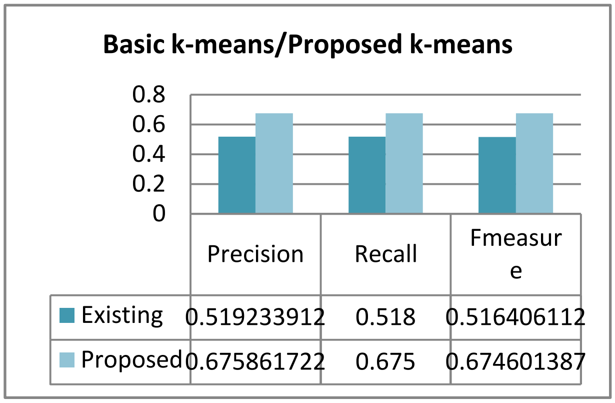
Dataset	Categories	No. of documents
Mininewsgroup20	alt.atheism comp.os.ms-windows.misc	100 100
Mini_dataset	Positive Negative	25 25
20_newsgroups	rec.sport.hockey rec.sport.baseball	500 500

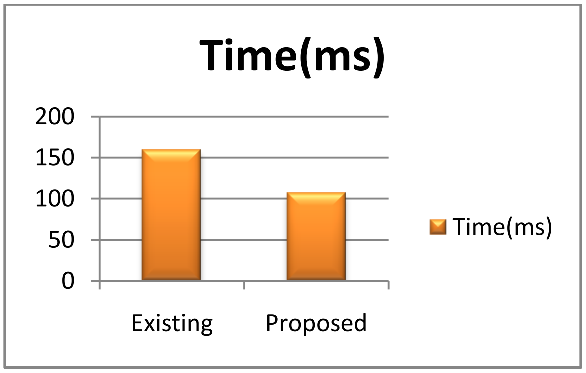
Performance evaluation based on precision, recall, F-measure and time.

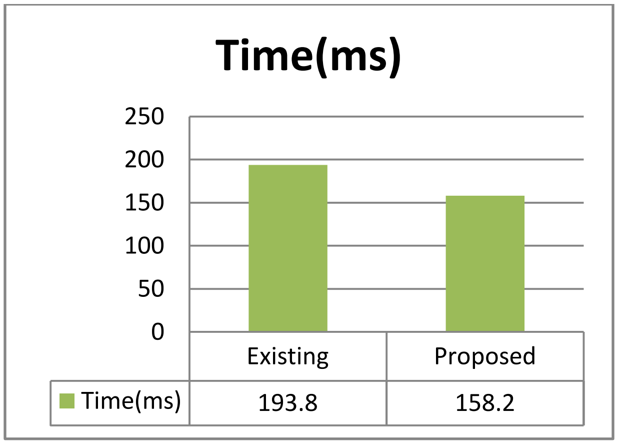
For Mininewsgroup Dataset:

	Precision	Recall	F-measure	Time ms)
Basic k-means	0.519234	0.518	0.516406	193.8
Proposed k-means	**0.675862**	**0.675**	**0.674601**	**158.2**







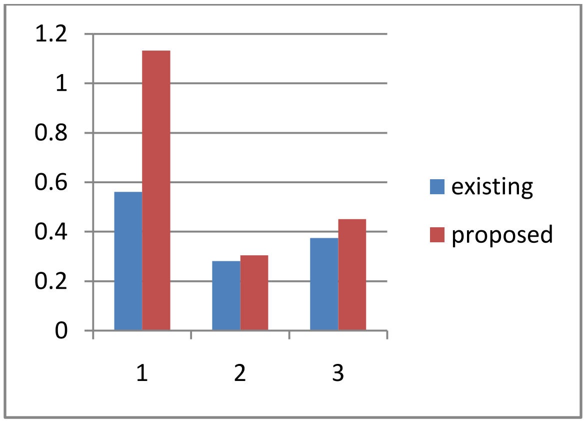


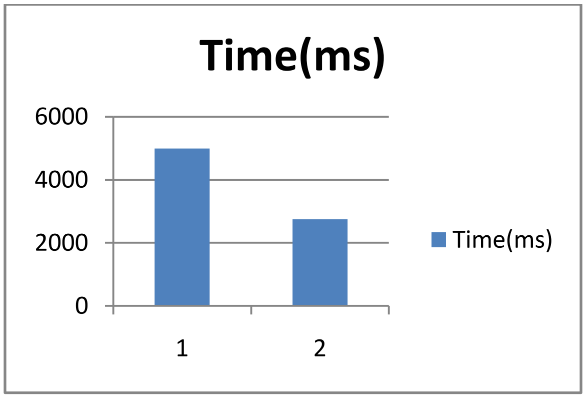
For 20_newsgroups Dataset:

	Precision	Recall	F-measure	Time (ms)
Basic k-means	0.561666	0.280 8	0.374325	4989. 2
Proposed k-means	1.132072	0.305	0.450853	2745. 8

For Mini_dataset:

	Precision	Recall	F-measure	Time (ms)
Basic k-means	0.540126	0.54	0.539639	158.8
Proposed k-means	**0.564004**	**0.564**	**0.563993**	**106.2**





6. Conclusion

In this paper, attribute weighted k-means for the clustering of documents is presented. It is shown that it is a promising mean for the document clustering as well as presenting the clusters in a meaningful way to the user. Experimental results shows that the proposed document clustering exhibits improvement in clustering of similar documents. F-measure is high for proposed algorithm and time required for clustering is less as compare to existing algorithm.

References

[1]Abhilash C B, Sharana Basavanagowda, “A Comparative Study on Clustering of Data using Improved K-means Algorithms”, *International Journal of Computer Trends and Technology (IJCTT)* , vol. 4, Issue 4, ISSN: 2231-2803, 2013

[2] Anna Huang, “Similarity Measures for Text Document Clustering”, Department of Computer Science

[3] Charu C. Aggarwal, Cheng Xiang Zhai, “A Survey of Text Clustering Algorithms”, IBM T. J. Watson Research Center Yorktown Heights

[4] Kemal Polat, “Application of Attribute Weighting Method Based on Clustering Centers to Discrimination of Linearly Non-Separable Medical Datasets”, Journal of Medical Systems, 2011

[5] Khaled M. Hammouda, Mohamed S. Kamel, “Efficient Phrase-Based Document Indexing for Web Document Clustering”, *IEEE Transactions On Knowledge And Data Engineering*, vol. 16, No. 10, 2004

[6] Marcelo N. Ribeiro, Manoel J. R. Neto, Ricardo B. C. Prudˆencio, “Local feature selection in text clustering”, Universidade Federal de Pernambuco

[7] Martin H.C. Law, Ma´ rio A.T. Figueiredo, Senior, Anil K. Jain, Fellow,” Simultaneous Feature Selection and Clustering Using Mixture Models”, *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, vol. 6, No. 9, 2004

[8] Miss. Vrinda khairnar , Miss. Sonal Patil,”Efficient clustering of data using improved K-means algorithm: A Review”, *Imperial Journal of Interdisciplinary Research (IJIR)*, vol.2, Issue-1 , ISSN : 2454-1362, 2016

[9] Neha Soni, Amit Ganatra, ”Categorization of Several Clustering Algorithms from Different Perspective: A Review*”, International Journal of Advanced Research in Computer Science and Software Engineering,* ISSN: 2277 128X, vol. 2, Issue 8, 2012

[10] Pranjal Singh, Mohit Sharma, “Text Document Clustering and Similarity Measures”, Dept. of Computer Science & Engg., 2013

[11] Salem Alelyani, Jiliang Tang, Huan Liu,” Feature Selection for Clustering: A Review”

[12] Samir Kunwar, “Text Documents Clustering using K-Means Algorithm”, 2013

[13] Sharmila, Shanthalakshmi Revathy.J, “An Efficient Clustering Algorithm for Spam Mail Detection”, *International Journal of Advanced Research in Computer Science and Software Engineering,* vol. 3, Issue 3, ISSN: 2277 128X, 2013

[14] Sujata Kolhe, Dr. Sudhir Sawarkar, “ Review of Document Clustering Techniques: Issues, Challenges and Feasible Avenue”, *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 5, Issue 4, ISSN: 2277 128X, 2015

[15] Sunita Bisht, Amit Paul, ”Document Clustering: A Review”, *International Journal of Computer Applications*, vol. 73, No.11, 2013

[16] Tatsunori Mori, Miwa Kikuchi, Kazufumi Yoshida, “Term Weighting Method based on Information Gain Ratio for Summarizing Documents retrieved by IR systems”, Div. of Electrical and Computer Eng.

[17] Wael M. S. Yafooz, Siti Z.Z. Abidin, Nasiroh, Omar, Rosenah A. Halim, ”Dynamic Semantic Textual Document Clustering Using Frequent Terms and Named Entity*”, IEEE*

3rd International Conference on System Engineering and Technology, 2013

[18] Yogesh Jain, Amit Kumar Nandanwar, "A Theoretical Study of Text Document Clustering", *International Journal of Computer Science and Information Technologies,* vol. 5(2), 2246-2251,ISSN: 0975 9646, 2014

[19]https://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html

[20]https://www.cs.cornell.edu/people/pabo/movie-review-data/