An Improved Algorithm for Text Document Clustering

Latika

Department of Computer Science and Applications
Kurukshetra University, Kurukshetra, India
latikavats353@gmail.com

Abstract- Due to the advancement of internet, the volume of the electronic documents available on the web is increasing day by day. Document clustering plays important role in organization and summarization of these documents. Thus, developing a fast and effective document clustering algorithm is of great importance. This paper presents an improved algorithm for document clustering. This algorithm is an enhancement of standard k-means algorithm. Experiments are conducted to evaluate the performance of improved algorithm and the results show that improved algorithm performs better than standard k-means algorithm. In this paper, feature selection is also applied to improve the clustering effectiveness.

Keywords- Document clustering, Improved k-means, Partitioning clustering, K-means, Feature selection, F-measure, VSM

I. INTRODUCTION

With the advancement of internet and less expensive hardware, there is a tremendous growth in the volume of electronic documents (news articles, scientific papers, blogs etc) available on the web. With web search engines a user can browse and locate the documents quickly but a web search engine often returns thousands of pages in response to a query, making it difficult for user to browse or to identify useful information. This issue has been widely studied using various techniques; document clustering is one of those techniques. Document clustering is a technique that is used to divide the documents into various groups called clusters such that documents in a group are similar to each other and objects in different groups are dissimilar. Document clustering can be used to automatically group the retrieved documents into a list of meaningful categories. By grouping similar documents together, a document collection can be easily browsed and queried [13]. Document clustering plays important role in document organization, summarization, and information retrieval.

Many clustering algorithms are available in literature for data clustering [6]. Out of these two major categories of clustering algorithms that are widely used for document clustering are: Partitioning clustering and Hierarchical clustering. Partitioning clustering divides the objects into fixed number of clusters which satisfy two requirements: 1) each group must contain at least one object 2) each object must belong to exactly one group [6]. Hierarchical clustering creates a hierarchical decomposition of the given documents. Hierarchical clustering can be divided into two categories: one is agglomerative and other is divisive. In agglomerative approach, also called bottom-up approach initially each object forms a separate cluster. It then successively merges the objects or clusters that are close to one another until a termination condition holds or until all clusters are merged into one. In divisive approach, also called top-down approach initially all the objects are in the same cluster. It then successively divides the cluster into smaller clusters until a termination condition holds or each object is in one cluster [6]. Partitioning clustering generally performs better than hierarchical clustering [8][16].

K-means is the most popular among all other partitioning clustering algorithms due to its simplicity and easy implementation. Although k-means is simple, but it is quite sensitive to the selection of initial centroids i.e. results of clustering depend on the initial positions of cluster centers. In standard k-means, the initial cluster centers are chosen randomly may cause solution to converge to local optimal. Several attempts have been made by researchers to replace this random initialization of initial cluster centers. This paper also presents an improved k-means algorithm in which initial centers are calculated instead of random selection.

This paper is organized as: Section II provides a brief overview of various approaches used for document clustering in last few years. Section III presents methodology used for document clustering. Experimental results are shown in Section IV. Finally, the conclusion of this paper is given in Section V.

II. LITERATURE REVIEW

Document clustering has been widely studied in computer science literature. Various approaches have been used for categorizing similar documents together. A brief overview of these approaches is given below.

Ying Zhao and George Karypis [16] performed comparison of six partitional and nine hierarchical agglomerative algorithms. The experimental results show that partitional algorithms produce better results than agglomerative methods.

Michael Steinbach et al. [8] performed comparison of k-means, bisecting k-means and UPGMA. Experimental results show that bisecting k-means is better than k-means and UPGMA and UPGMA is best among three agglomerative techniques.

Hsi-Cheng Chang et al. [3] proposed a document clustering approach based on keyword cluster method. This method has mainly two benefits: first, it greatly reduces the computing complexity and another, it provides higher classification accuracy. The proposed method has been compared with single link, complete link, average link and k-means algorithms and precision of proposed method is found to be highest among all.

Mushfeq-Us-Saleheen Shameem and Raihana Ferdous [10] proposed an efficient k-means algorithm for document clustering. In the proposed method Jaccard distance measure is used with k-means algorithm for finding most dissimilar k documents for centroids of k clusters.

Yi-hong Lu and Yan Huang [17] proposed an entropy based TF/IDF classifier which uses probabilistic indexing paradigm. The proposed algorithm uses Rocchio Algorithm with TF/IDF weighting scheme.

O.H Odukoya et al. [12] proposed an improved algorithm for document clustering. The proposed algorithm is a variation of k-means algorithm that introduces a new initialization method for selection of initial centers for k-means algorithm.

Muhammad Rafi et al. [9] presented a novel approach for document clustering based on topic maps. Topic map representation of documents captures the semantics of documents. The knowledge extracted from topic maps is used to measure similarity between documents.

M. Arshad [11] proposed a new clustering algorithm Kea bisecting K-means which is baesd on KEA and Bisecting k-means algorithm. The proposed method uses KEA (Automatic Key-phrase Extraction) for extracting several key phrases from documents and Bisecting k-means algorithm for performing clustering on these extracted key phrases.

Fathi H. Saad et al. [2] performed comparison of various hierarchical agglomerative algorithms. The comparison of algorithms has included six criterion functions that can be divided into four groups (internal, external, graph based and hybrid) and three selection schemes (single-link, complete link and UPGMA). Cosine similarity measure has been used for measuring similarity.

Leena. H. Patil and Mohammed Atique [7] presented a novel approach for feature selection method tf-idf. Three different term weighting schemes tf-idf, tf-df, and tf 2 have been used and the terms having weight higher than a pre-specified threshold are selected as key features. Tf-idf feature selection scheme is found to be better among all.

Sunghae Jun et al. [15] presented a method to overcome the problem of sparseness in document clustering. To remove sparsity in document term matrix SVD-PCA is used. K-means clustering based on support vector clustering and Silhouette measure is used for performing clustering operation.

Yinglong Ma et al. [18] presented a three phase approach to document clustering. In first phase the best topic model is determined, in second phase initial clustering centers are obtained using k-means++ and in last phase k-means is applied for document clustering.

III. METHODOLOGY

Clustering algorithms cannot be directly applied to documents. Before applying clustering algorithms, several steps need to be followed. A brief description of these steps is given below.

Pre-processing

The input to this step is a plain text document and the output is a set of pre-processed tokens. For document pre-processing following steps are followed.

- A document consists of several sentences. Firstly, words are extracted from these sentences. These words are called tokens.
- All the special symbols, punctuations and digits are removed.
- Upper case letters are converted to lower case letters.
- All the stop words are removed. Stop words are those frequently words which are insignificant because they do not carry any meaning from clustering point of view. For e.g. on, over, so etc. For stop words

removal, a list of stop words is created and then each token is compared with this list. If token is a stop word then it is removed else it is kept for further processing.

- Word stemming is performed. In word stemming, all the words are reduced to their stem word. For e.g. all words like 'looks', 'looked' and 'looking' are reduced to their stem word 'look'. For performing stemming of words Porter's algorithm is used.

Feature Selection

After pre-processing, all the unique words are extracted. Feature Selection is applied to these unique words. Feature selection removes all the features that do not have any discriminating features. For feature selection, a threshold value is defined and then, all the words having frequency less than threshold value are neglected and all the words having frequency greater than the threshold value are selected. Feature selection returns frequent and relevant features for each category. Feature selection results in dimension reduction and also increases the clustering effectiveness.

Document Representation

For clustering documents, they need to be represented in document representation model. Vector Space Model (VSM) is a widely used document representation model. Under VSM, n documents with m unique words are represented as n×m document-term matrix that is each document is a vector of m dimension. Each document has different weights for different words as shown in figure 1.

	Word 1	Word 2	. .	Word m
Document 1	$W_{1,1}$	$W_{1,2}$		$W_{1,m}$
Document 2	$W_{2,1}$	$W_{2,2}$		$W_{2,m}$
.				
Document n	$W_{n,1}$	$W_{n,2}$		$W_{n,m}$

Figure 1.Document by word matrix

Term frequency (TF) is used for finding out weights of each word in each document. Term frequency is how many times a term occurs in a document. It is a measure of significance of a term in a document. Assumption is that more frequently used words are significant. Under TF weights are represented as:

$W_{ij} = f_{ij}$

Where f_{ij} : Frequency of jth word in ith document.

Clustering Algorithm

Clustering algorithm organizes similar documents into clusters. Document clustering algorithms take VSM as input and provide clusters of documents as output. In this work, two clustering algorithms are used for performing clustering of documents. One is basic k-means algorithm and another is improved k-means algorithm.

Basic K-means Algorithm

K-means is widely used for document clustering. The basic steps involved in k-means are shown in figure 2. Here, n data vectors represent n document vectors of VSM.

Input:

D: a set of n data vectors

k: number of clusters

Output:
A set containing k clusters

Steps:

1. Arbitrarily choose k data vectors from n for determining initial cluster centers and assign them to c_j ($1 \leq j \leq k$)
2. Calculate the distance of each data vector in D to each cluster center in C.
3. Assign each data vector to the cluster to which the data vector is most similar i.e. having minimum distance.
4. Update the cluster centers by calculating the mean of data vectors assigned to a cluster.
5. Repeat steps 2 to 4 until stopping criteria is met.

Stopping criteria:

Number of iterations or change in the position of cluster centers in consecutive iterations.

Figure 2.Steps involved in basic k-means

Although, k-means is simple and basic algorithm for cluster analysis but the main problem with k-means is its random initialization method. Random initialization of initial cluster centers may cause solution to converge to local optimal. Due to this reason the random initialization method of k-means is replaced in improved k-means algorithm.

Improved k-means

Improved k-means is an enhancement of standard k-means algorithm. This enhanced method is taken from the method proposed by [1]. In this algorithm, initial cluster centers are selected by performing some calculation. The whole process of improved algorithm is divided into two phase.

Phase 1: The input to this phase is n document vectors of VSM. Out of these vectors initially k clusters are formed by dividing vectors into k sub- arrays. Size of these clusters is fixed.

Phase 2: In second phase, initial clusters obtained from first phase are passed as input. Initial cluster centers are calculated by taking average of these clusters. After that final clusters are computed. Various steps involved in improved k-means are shown in figure 3.

Input:

D: a set of n data vectors

k: number of clusters

Output:

A set containing k clusters

Steps:

Phase 1:

1. Determine the size S of cluster by using round (n/k).
2. Create k arrays C_i ($1 \leq i \leq k$)
3. Transfer S data vectors from input set D to each C_i until all data vectors are removed from D.
4. k initial clusters are obtained.

Phase 2:

1. Compute the average Av_i of each cluster C_i ($1 \leq i \leq k$) to determine initial cluster centers.
2. Find out distance Dt of each data vector in cluster C_i from its own cluster center Av_i ($1 \leq i \leq k$).
3. Find out the distance Ds of each data vector in V from each cluster center Av_i ($1 \leq i \leq k$).
4. If Dt $\leq$ min (Ds) then data vector stay in same cluster else data vector will move to that cluster to which it is most similar or having less distance.
5. Recalculate the cluster centers & move data vectors until no change.

Figure 3.Steps involved in improved k-means

Evaluation: The performance of both algorithms is measure in terms of F-measure, Time complexity and Number of iterations.

F-measure: F-measure is harmonic mean of precision and recall. Precision can be defined as ratio of number of relevant documents to the number of total documents retrieved. Recall can be defined as the ratio of number of relevant documents retrieved to the total number of relevant documents. Consider a class i and cluster j, then

$$\text{Prec}\ (i,j) = \frac{n\ (i,j)}{\text{nj}}$$

$$\text{Recall}\ (i,j) = \frac{n\ (i,j)}{\text{ni}}$$

Where n (i, j) represents number of documents of class i present in cluster j. ni represents number of documents in class i. nj represents number of documents in cluster j. Then, F-measure of class i with respect to cluster j is given as:

$$\text{F}\ (i,j) = \frac{2 * \text{prec(i, j)} * \text{recall(i, j)}}{(\text{prec(i, j)} + \text{recall(i, j)}}$$

The average F-measure of is given as:

$$\text{F} = \frac{1}{\text{k}} \sum_{i=1}^{\text{k}} \text{F}\ (i,j)\ \ | 1 \leq \text{j} \leq \text{k}\}$$

IV. EXPERIMENTAL RESULTS

All the experiments are performed using MATLAB 7.8.0 on a laptop computer with following key configuration: window vista, Intel (R) Pentium (R) Dual CPU @ 2.16 GHz, and 2 GB RAM. Both the clustering algorithms are evaluated by using mini_newsgroups dataset [4]. This dataset contains 20 newsgropus categories where each category contains 100 documents. This experiment is performed on the three categories of this dataset: alt.atheism, comp.graphics, and comp.os.ms-windows.misc.

A list of 456 stop words is used for stop words removal. Porter's algorithm that is used for performing stemming of words is taken from [5]. The experiment is conducted several times for three categories of mini_newsgroups dataset. In each experiment, the results are computed and then the average value of these results is taken as final results. Table 1 shows the final results obtained for both clustering algorithms.

Table I. Final results obtained for three solutions

	F-measure	No. of iteration	Time comp.
K-means	0.1720	11	0.2587
Improved k-means	0.5175	6	0.1641

The graphical representation of comparisons between both clustering algorithms on the basis of different evaluations measures is shown in the following figures.

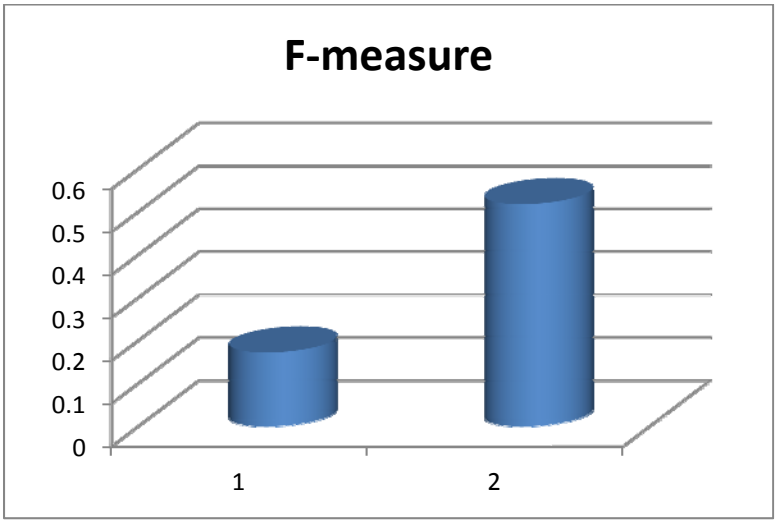

Figure 4.F-measure comparison between k-means and improved k-means

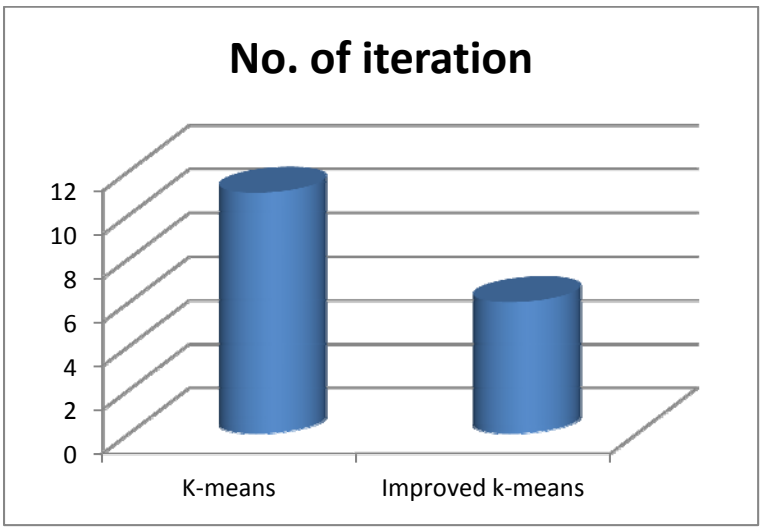

Figure 5.No. of iteration comparison between k-means and improved k-means

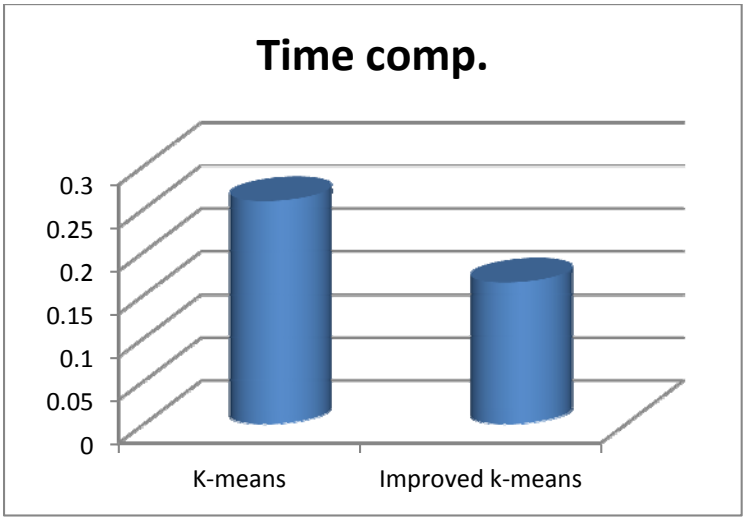

Figure 6.Time complexity comparison between k-means and improved k-means

V. CONCLUSION

In this paper, an improved k-means algorithm is used for performing clustering on documents. To measure the performance of improved algorithm, it is compared with Standard k-means. The experimental results show that the improved k-means clustering algorithm outperforms the standard k-means algorithms in terms of f-measure, No of iterations and time complexity.

REFERENCES

[1] Azhar Rauf, Sheeba, Saeed Mahfooz, Shah Khusro and Huma Javed, "Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity", Middle-East Journal of Scientific Research, Vol.12, No.7, pp. 959-963, 2012.
[2] Fathi H. Saad, Omer I. E. Mohamed and Rafa E. Al-Qutaish, "Comparison of Hierarchical Agglomerative Algorithms for Clustering Medical Documents", International Journal of Software Engineering & Applications, Vol.3, No.3, May 2012.
[3] Hsi-Cheng Chang, Chiun-Chieh Hsu and Yi-Wen Deng, "Unsupervised Document Clustering Based on Keyword Clusters", International Symposium on Communications and Information Technologies 2004, Oct. 2004.
[4] https://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html
[5] http://tartarus.org/martin/PorterStemmer/matlab.txt
[6] J. Han and M. Kamber, "Data Mining concepts and techniques", Morgan Kaufmann Publishers, Second edition, 2009.
[7] Leena. H. Patil and Mohammed Atique, "A Novel Approach for Feature Selection Method TF-IDF in Document Clustering", 3rd IEEE International Advance Computing Conference (IACC), 2013.
[8] Michael Steinbach, George Karypis, and Vipin Kumar, "A comparison of document clustering techniques", In KDD Workshop on Text Mining, 2002.
[9] Muhammad Rafi, M. Shahid Shaikh and Amir Farooq, "Document Clustering based on Topic Maps", International Journal of Computer Applications (0975 – 8887) , Vol.12, No.1, Dec. 2010.
[10] Mushfeq-Us-Saleheen Shameem and Raihana Ferdous, "An efficient K-Means Algorithm integrated with Jaccard Distance Measure for Document Clustering", IEEE, 2009.
[11] M. Arshad, "Implementation of Kea-Keyphrase Extraction Algorithm By Using Bisecting k-means Clustering Technique For Large And Dynamic Data Set", International Journal of Advanced Technology & Engineering Research, Vol.2, Issue 2, Mar.2012.
[12] O.H Odukoya, G.A. Aderounmu and E.R. Adagunodo, "An Improved Data Clustering Algorithm for Mining Web Documents", IEEE, 2010.
[13] Ramiz M. Aliguliyev, "Clustering of document collection – A weighting approach", Expert Systems with Applications 36, pp. 7904-7916, 2009.
[14] Ran Vijay Singh and M.P.S Bhatiya, "Data Clustering with modified K-means Algorithm", IEEE- International Conference on Recent Trends in Information Technology, June 2011.
[15] Sunghae Jun, Sang-Sung Park and Dong-Sik Jang, "Document clustering method using dimension reduction and support vector clustering to overcome sparseness", Expert Systems with Applications 41, pp. 3204–3212, 2014.
[16] Ying Zhao and George Karypis, "Evaluation of Hierarchical Clustering Algorithms for Document Datasets", Technical Report, Jun. 2002.
[17] Yi-hong Lu and Yan Huang, "Document Categorization with Entropy based TF/IDF classifier", IEEE, 2009.
[18] Yinglong Ma, Yao Wang and Beihong Jin, "A three-phase approach to document clustering based on topic significance degree", Expert Systems with Applications 41, pp. 8203–8210, 2014.