Text Document Categorization using Modified K-Means Clustering Algorithm

Sheelesh Kumar Sharma, Navel Kishor Sharma

Abstract: ***The volume of the information that is to be managed is increasing at exponential pace. The challenge arises how to manage this large data effectively. There are many parameters on which the performance of such a system can be measured such as time to retrieve the data, similarity of documents placed in same cluster etc. The paper presents an approach for auto-document categorization using a modified k-means. The proposed methodology has been tested on three different data sets. Experimental findings suggest that proposed methodology is accurate and robust for creating accurate clusters of documents. The proposed methodology uses cosine similarity measure and a fuzzy k-means clustering approach to yield the results very fast and accurately.***

Index Terms: K-means, Text mining, Web mining

I. INTRODUCTION

The effective management of huge text is a challenge. The mammoth size of text is available on Internet so as the on the cloud [16-18], in the form of web pages, blogs, news articles, social media interactions, business interactions etc. Clustering is a type of unsupervised machine learning technique where the labels of the classes are not known in advance. The clusters are defined out of the data, based on certain similarity measures. By creating clusters of documents, we may reduce the large number of documents into much smaller number of clusters which are coherent. Here, coherence means that documents in a particular cluster are similar to other documents in same cluster and are different from the documents in other clusters.

The area of text mining has been a hot area of research for quite some time. The clustering approaches can be roughly classified into four categories namely flat, hierarchical, hard and soft clustering. K-Means clustering falls under the category of flat clustering, where clusters do not fall under any specific structure. Allahyari et al.[1] gives a nice survey on various approaches for text mining including many popular classification and clustering approaches for text mining. Same has been presented in [2].

II. LITERATURE SURVEY

K-Means clustering was first used in year 1955 and since then thousands of the algorithms have been proposed but still this algorithm is widely used [3]. It speaks volumes about the difficulty in devising a general algorithm for clustering. For text based applications, k-means still happens to be the favorite choice even today. Variants of k-means can be found in the literature. An entropy weighted k-means algorithm for high dimensional sparse data has been presented in [4], which can generate better clustering results than other subspace clustering algorithms.

Similarity or distance (dissimilarity) measures play an important role in a clustering algorithm. A similarity or distance measure is something that decides whether a document is to be placed in cluster x or in cluster y based on how similar or dissimilar it is from other documents in that cluster. There are various similarity measures available such as cosine similarity, Jaccard correlation coefficient etc and there are many distance measures such as Euclidean distance, Mahalanobis distance, relative entropy distance etc [5] [6].

III. PROPOSED METHODOLOGY

The proposed approach is divided into multiple steps including per-processing, document representation, feature extraction and clustering. The approach has been tested on three different datasets [7]-[9].

A. *Pre-processing Step*

As a first step of the proposed methodology, the text need to be pre-processed as follows:

a. Eliminate the common terms from the documents. We have written a simple Python script which eliminates very common terms from each document (such as helping verbs, articles, prepositions -- "is", "am", "then"). We have created a corpus of such common words that are used for elimination from all the data sets used for testing.

b. Convert different forms of a word into single canonical form e.g words "report", "reporting", "reporting", "reports" need to replaced with the single instance of word "report".

B. *Document Representation*

Each document is represented in the form of a vector of words and this model is called as document space vector model. Many available techniques for text categorization used term or word frequency but our experimental finding suggest that instead of using term frequency we may use the square root of it, which happens to be a more precise metric for further training.

The square root metric tend to improve the purity and entropy of the categorization results and avoids skewedness towards mean error rate. The representation of each document is given by a vector:

Revised Manuscript Received on July 5, 2019.

Dr Sheelesh Kumar Sharma, Professor (Comp. Sc.), IMS Ghaziabad, India,

Mr Navel Kishor Sharma, **Navel Kishor Sharma, Associate Dean,** Academic City College Ghana

Word Frequency Vector WFV= (sqwv$_1$, sqwv$_1$, sqwv$_2$,.....sqwv$_n$) (1)

Where sqwv$_i$ is the frequency of the ith word in the document. Similar to word frequency, the term frequency (tf) may be defined as follows:

$$tf(t,d) = 0.5 + 0.5xf_{t',d}\,(maxf'_{t',d} : t' \in d) \quad (2)$$

There is another associated measure called inverse document frequency (IDF). It refers to the weightage of a word in providing unique useful information. If a word is common across the documents, then it will not have much of the weightage. Inverse document frequency is calculated by

IDF= log (total number of documents / number of documents containing a particular term)

$$idf(t,d) = log\frac{N}{[mode(d \in D : t \in D)]} \quad (3)$$

IDF is the measure whether term is a unique characteristic of document or it is common across all the documents.

C. *Similarity Measure*

For creating clusters, various similarity measures exist. Choice of the right similarity measure may greatly affect the clustering accuracy and performance. For our document categorization task, we evaluated three similarity measures namely Euclidean Distance, Jaccard Coefficient and Cosine Similarity. Cosine Similarity is the best performing measure out of all the three.

D. *Modified K-Means*

Here, we use a modified k-means algorithm that we call as fuzzy k-means algorithm. The idea is that instead of saying that a document point belongs to some specific cluster, it belongs to all clusters. This belongingness is defined by membership function.

Step 1: Estimate the number of clusters K

Step 2: Initialize the k seeds for the centroid of the clusters.

Step 3: Use the cosine similarity measure to find the centroid of the clusters. The Cosine Similarity of two documents doc$_1$ and doc$_2$ can be given as follows:

$$cosine(doc_1, doc_2) = (doc_1 . doc_2)/[length(doc_1)Xlength(doc_2)] \quad (4)$$

Step 4: Calculate the distance of all document points from centroids and assign them new clusters as algorithm proceeds. The membership of a document may range from 0 to 1.

Step 5: Repeat the step until all memberships converge to 0 or 1 or a specified number of iterations have been completed.

E. *Knowing the Ideal Value of Number of Clusters K*

K-means approach is a flat clustering approach. Let K be the number of desired clusters, then it finds K clusters at once. When compared with traditional hierarchical approaches of clustering, which divide or join clusters to get new set of clusters, the time complexity of k-means is much less (almost linear).

Knowing the best value of number of clusters for a clustering approach is a big challenge. In literature, researchers have suggested a few techniques that can provide a rough idea on the appropriate value for K [10]-[12]. Most of these methods use hit and try approach for finding the value of K. An accurate estimation of the value of K helps in achieving better accuracy of clustering.

IV. Experimental Set-up and Different Text Document Datasets

Document representation: Each document is represented using vector model. Here, we use a vector space model which uses TF-IDF (term frequency – inverse document frequency) model. Term frequency refers to the number of times a particular term repeats in the document. The term frequency of a word is calculated by total occurrences of the word divided by total words in document.

For evaluation of the proposed methodology, three bench-marked datasets have been used. These datasets are 20 Newsgroups Dataset [7], Classic 4 Dataset [8], and CMU Knowledge-base Dataset [9]. These datasets represent three different domains of documents. 20 Newsgroups Dataset has the collection of 18828 documents of 20 categories. Classic Dataset is the collection of research papers from 4 different disciplines. Lastly, CMU Web Knowledge-base dataset is the collection of 8282 web pages representing 7 different categories. The average clustering accuracy of the proposed approach on above data sets is 96.56%. Purity and entropy are the two measures that are used for evaluating an approach for a clustering algorithm. We have tested the proposed methodology on three different data sets. There is a practical challenge for evaluating the clustering algorithms. It is the unavailability of hand-labeled or annotated data. In such a scenario, other performance measures become handy in giving an idea of the performance of the system.

Table 1: Description of the Datasets used for Evaluation of Proposed Methodology

S.No	Dataset	No of Docs	Categories	Terms	Avg Class Size	Type of Documents
1	20 News Groups Dataset [7]	18828	20	28553	1217	Newsgroup post data
2	Classic Dataset [8]	7095	4	12009	1774	Academic papers falling under 4 categories namely- **CACM**: 3204 documents **CISI**: 1460 documents **CRAN**: 1398 documents **MED**: 1033 documents
3	CMU Web Knowledge Base Dataset[9]	8282	7	20682	1050	Collection of web pages

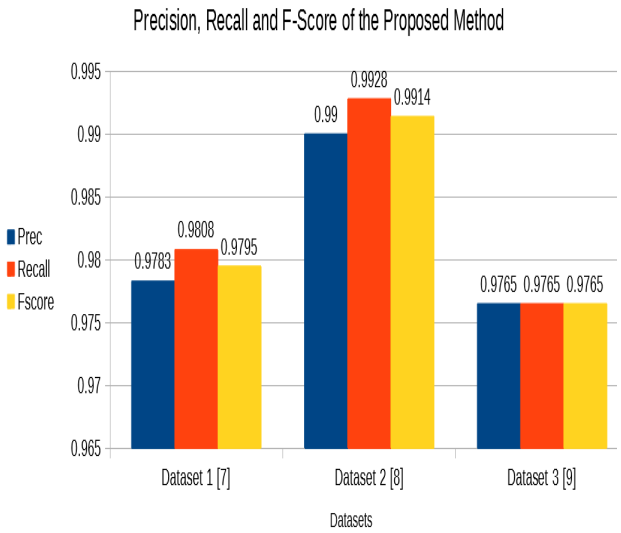

Fig. 1: Performance of proposed method

The mostly used metrics are between-cluster distances and within-cluster distances. However, manually assigned labels are normally not available in clustering, and in these cases other measure such as within-cluster distances and between-cluster distances [6] can be used for evaluation. These are not used in this paper because all the data sets already have labels.

I. Results and Evaluation of Proposed Method on Various Performance Parameters

Various measures exist for evaluating the performance of a clustering technique such as precision, recall, f-score, entropy etc. The viability of the proposed technique has been reported in terms of the three measures namely precision, recall and f-score.

Before further discussion, lets understand some basic terms associated with errors in clustering. These are false positives (FP) and false negatives (FN). If two dissimilar documents are kept in one cluster then it will be an instance of false positives. Likewise, if two similar documents are kept in different clusters, then it will be an instance of false negative. Let TP be the number of total positive instances of documents, then the accuracy measures can be calculates as follows:

$$Precision\ (P) = TP/\ (TP + FP)$$
$$Recall\ (R) = TP/\ (TP + FN)$$

Once the values of P and R are calculated, then the value of f-score can be easily calculated. F-score is just the harmonic mean of precision and recall of the clustering system.

$$F\text{-}Score\ (FS) = 2\ x\ (P\ x\ R)/\ (P + R)$$

Table 2 presents these measures on all the three different data sets used to evaluate the proposed method of document clustering.

Fig. 1 depicts the pictorial representation of the three performance parameters for the proposed method. Is is evident from the results that the method is quite suitable for clustering of the documents of different types. The average accuracy of proposed method on these three data sets happens to be 96.56%. The method gives best clustering results with Classic Dataset [8] and the least accuracy with CMU Web Dataset [9].

II. Conclusion

Categorization of text documents in suitable categories is an import task due to its various applications such as online content management, SEO, news reporting etc. The paper presents an approach for document categorization and used a modified k-means.

The viability of a clustering algorithm depends on the type of similarity measure. The similarity measure can be distance based or the concept based. This paper presents an approach for clustering text documents using a modified k-means algorithm. The proposed technique is tested on three different datasets. Experimental results show that proposed approach is accurate and robust for creating clusters of documents.

Table 2: Performance Measures of Proposed Method

Datasets	Total Docs	Rightly Classified	Mis-classified	Accuracy(%)	TP	FP	FN	Precision	Recall	F-Score
Dataset 1 [7]	18828	18075	753	96	18075	400	353	0.9783	0.9808	0.9795
Dataset 2 [8]	7095	6975	120	98.3	6975	70	50	0.99	0.99	0.9900
Dataset 3 [9]	8282	7902	380	95.4	7902	190	190	0.98	0.98	0.9800

REFERENCES

1. Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saied Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys Kochut. 2017. A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques . In Proceedings of KDD Bigdas, Halifax, Canada, August 2017, 13-26.
2. Berry, M. W. (2004). Survey of text mining. Computing Reviews, 45(9), 548.
3. Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern recognition letters, 31(8), 651-666.
4. M. K. Ng, J. Z. Huang and L. Jing, (2007) "An Entropy Weighting k-Means Algorithm for Subspace Clustering of High-Dimensional Sparse Data," in IEEE Transactions on Knowledge & Data Engineering, vol. 19, no. , pp. 1026-1041.
5. Huang, A. (2008, April). Similarity measures for text document clustering. In Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand (Vol. 4, pp. 9-56).
6. Neuhaus, J. M. and Kalbfleisch. J. D. (1998). Between- and within-cluster covariate effects in the analysis of clustered data. Biometrics, 54(2), pp 638–645
7. 20 News Net Dataset: http://qwone.com/~jason/20Newsgroups/ (last visited on February 15, 2019)
8. Classic dataset http://www.dataminingresearch.com/download/dataset/classicdocs.rar (last visited on February 15, 2019)
9. CMU Web Knowledgebase Dataset http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/ (last visited on February 15, 2019)
10. Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411-423.
11. Rasson, J. P., & Kubushishi, T. (1994). The gap test: an optimal method for determining the number of natural classes in cluster analysis. In *New approaches in classification and data analysis* (pp. 186-193). Springer, Berlin, Heidelberg.
12. Pham, D. T., Dimov, S. S., & Nguyen, C. D. (2005). Selection of K in K-means clustering. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, *219*(1), 103-119.
13. Revanasiddappa, M. B., & Harish, B. S. (2018). A New Feature Selection Method based on Intuitionistic Fuzzy Entropy to Categorize Text Documents. *International Journal of Interactive Multimedia & Artificial Intelligence*, *5*(3).
14. Nasser, S., Sreejith, C., & Irshad, M. (2018, July). Convolutional Neural Network with Word Embedding Based Approach for Resume Classification. In *2018 International Conference on Emerging Trends and Innovations In Engineering And Technological Research (ICETIETR)* (pp. 1-6). IEEE.
15. Lata, S., & Loar, M. R. (2018). Text Clustering and Classification Techniques-A Review. *International Journal on Recent and Innovation Trends in Computing and Communication*, *6*(3), 237-241.
16. *S.Kumar, R. Bahsoon, T. Chen, and R. Buyya. Identifying and Estimating Technical Debt for Service Composition in SaaS Cloud.25th IEEE International Conference on Web Services, 2019.*
17. *S.Kumar, R. Bahsoon, T. Chen, K. Li, and R. Buyya. Multi-Tenant Cloud Service Composition Using Evolutionary Optimization. 24th IEEE International Conference on Parallel and Distributed Systems, 2018.*
18. *A.Vashishtha, S. Kumar, P. Verma, and R. Porwal. A Self-Adaptive View on Resource Management in Cloud Data Center.8th International Conference on Cloud Computing, Data Science & Engineering, 2018*

AUTHORS PROFILE

Dr Sheelesh Kumar Sharma
He is a Professor in IT Department at IMS Ghaziabad. He obtained **MCA** from M. B. M. Engineering College Jodhpur, **M. Tech.** in Computer Science from Institution of Electronics and Telecommunication Engineers New Delhi and **M. Phil.** in Computer Science from Madurai Kamraj University Madurai. He has awarded **Ph.D.** on Web Based Learning from **University of Rajasthan**, Jaipur and his research area is data mining and data warehousing. He has more than 18 years of experience in academics and 6 years of Industrial experience. He is a life time member of the Institution of Electronics and Telecommunication Engineers (IETE) Delhi.

Navel Kishore Sharma is an Associate Dean at Academic City College Accra in Ghana. He did his M. Tech. from Rajasthan University, Jaipur in 2006 and pursuing PhD. He has more than **25 years** of professional experience in the fields of Software Development and Teaching in Engineering Colleges.