International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-2, December, 2019

Research of Clustering Algorithms using Enhanced Feature Selection

Venkata Nagaraju Thatha, A.Sudhir Babu, D.Haritha

ABSTRACT---In Present situation, a huge quantity of data is recorded in variety of forms like text, image, video, and audio and is estimated to enhance in future. The major tasks related to text are entity extraction, information extraction, entity relation modeling, document summarization are performed by using text mining. This paper main focus is on document clustering, a sub task of text mining and to measure the performance of different clustering techniques. In this paper we are using an enhanced features selection for clustering of text documents to prove that it produces better results compared to traditional feature selection.

Keywords: enhanced feature selection, text mining, clustering.

I. INTRODUCTION

In the current digital world, huge quantity of data is stored in variety of forms like image, text, audio, video and this may increase in the future[1]. The increase of digital information increases the demand of tools for analysis and discovers useful information. Text mining is a sub field of data mining used for analysis of documents[2]. The various tasks related to text are concept extraction, information extraction, document summarization, entity relation modeling and clustering . A sub task of text mining is document clustering, where documents are grouped into meaningful clusters such that the documents are similar to each other with in the cluster and dissimilar to other in different clusters. In text mining one of the major areas is clustering. It provides high level view for large amount of data to determine the relationship among the texts and arranges the text documents into valid clusters such that improve the similarity with in cluster and reduce similarity between different clusters[3]. Mostly the search engines, digital libraries use clustering of text documents. Document clustering is widely used in different areas like applications related to security, biomedical applications, online media, software applications, market applications, sentimental analysis and academic applications. The common aim of all these techniques is to extract information of high quality from the text. So, main aim of industry is to identify techniques that will enhance the discovery of knowledge.

The aim of this paper is to improve the performance of different clustering algorithms with the help of enhanced feature selection. Many number of techniques for text document clustering is proposed by several researchers[4]. In all these techniques of clustering the procedure to be followed are preprocessing, feature selection, dimensionality reduction and clustering algorithm[5]. Feature selection is a technique of determining the terms that are having greater impact on performance of clustering by removing unnecessary and irrelevant data .The selected features are generally high dimensional, which have more impact on performance of clustering algorithm. So, high dimensionality is to be reduced by the clustering algorithms by maintaining meaningful structures to the documents[6].

II RELATED WORK

A. *Data Preprocessing*

First data corpus is applied to preprocessing. In preprocessing the first phase is bag of words. The technique is identify the terms in the corpus and specify the count of each term that appeared in the document. This model does not consider the order in which the term appears and semantics[7]. After completion of Bag of words, Stop word removal is applied to the data corpus. In this unnecessary terms are removed from the data corpus. The stop word list contains the commonly used terms;

in',’a’,’the’,’for’,’since’,’on’,’between’ etc.

Now Stemming is performed on the data corpus. Stemming is the process of converting different words which have same prefix into root form. For example the words implementation, implementing, implementable etc are converted into implement[8]. Now the preprocessed data is applied to feature selection[9].

B. *Traditional Feature Selection*

The traditional feature selection is

$$Weight_{ij} = TF\text{-}R(term,doc) * IDF\text{-}R(term) \quad (1)$$

Where

$$TF\text{-}R(term,doc) = 0.5 + .5 * f_{term,doc} / \max\{f_{term,doc} : t \in d\} \quad (2)$$

is used to compute the count of a term present in the given document.

$$IDF\text{-}R(term) = \log(N/n) \quad (3)$$

is the overall documents per count of the term that present within the given document.

Next, the preprocessed data is applied to enhanced feature selection to increase the performance of clustering algorithms.

Revised Manuscript Received on December 30, 2019.
* Correspondence Author
VenkataNagarajuThatha, Department of Computer Science & Engineering, JNTUK UNIVERSITY, Kakinada, Andhra Pradesh India. (Email: nagarajuthatha@gmail.com)
A. SudhirBabu, Department of Computer Science & Engineering, PVPSIT, Vijayawada, Andhra Pradesh India. Email: asbabu@hotmail.com
D. Haritha, Department of Computer Science & Engineering, JNTUK UNIVERSITY, Kakinada, Andhra Pradesh India. (Email: harithaphd1@gmail.com)

C. *Enhanced feature selection*

The Enhanced feature selection is

$Weight_{ij}$= TF-R(term,doc)* IDF-R(term)* K_1*K_2 (4)

Where K_1 specifies significant of terms calculated as TF-r/IDF-R and K_2 is the distinguish ability of the terms.

III PROPOSED SYSTEM

A. *K-Means Clustering Algorithm*

k-means is the most commonly used partition clustering algorithm for document clustering. It clusters the documents into different groups such that each cluster maintains similarity among the documents. It works as assign the points to nearest cluster based on the centroid. Cluster centroid is calculated by taking average for all the data points present within the cluster. k-means algorithm mainly depends on how we assign points to clusters. At every stage of algorithm, the data points are allotted to the closet partition based on similarity measures like cosine similarity or Euclidean distance[10] etc. At each stage the partitions are recalculated such that points are switched from one partition to another partition.

The steps to be followed are:

- Assume k - number of clusters.
- Compute the centroids for k clusters.
- Allot every point to their nearest centroid of the cluster.
- After assignment again evaluate centroids of the clusters.
- Repeat the above steps until centroid of the cluster is fixed.
- For k-means algorithm the input is of partitions (clusters). Generally, k – means algorithm produces k distinct clusters. The advantage of this algorithm is, simple and applicable to large data sets[11].

B. *EXPECTATION MAXIMIZATION*

Generally, by using EM algorithm maximum likelihood parameters are identified for a statistical model.nce the variables or parameters are identified they are assigned to clusters. The basic step is a large continuous variable in a large group of observations are computed[12]. The dataset contains two clusters of variables with each cluster having different mean and by using normal distribution distributes the values of large continuous variable. The EM algorithm extend the approach of clustering in a way such that: Instead of normal assignment of variables to the clusters that increases difference in means, the EM algorithm cluster membership probabilities are calculated by using some probability measurements. The EM algorithm is applied to categorical and c continuous variables. EM algorithm is divided into two steps. i) Expectation (E)-step ii) Maximization (M)- step. In the first step we estimate the missing data and in the second step maximize the likelihood function[13].

The process for effective clustering by using EM algorithm is,

- Collect required training data from the data corpus.
- Identify the missing values from the data corpus through iterative processing of training data.

Apply E-step and perform optimization of E-step by using the equation

$P(\mu,\mu(w))=\log Li(\mu;A)+E[\log Li(\mu,B)]$ (5)

Where observed data is A , missing data is B. the log Li is calculated by using conditional mean based on A,B and current estimate of μ.

Repeat he procedure until it identifies the unobserved information.

Apply M-Step. Compute maximization E-step by using equation (above) based on the M-step.

$\mu(w+1)$=paramaximum $P(\mu,\mu(w))$ (6)

It is observed that application of EM is simple .the reason is M-step is executed iteratively. The result of convergence identifies maximum likelihood function in increasing order but it does not produce a global solution in clustering of text documents even if it contains missing data. Finally EM algorithm needs more number of iterations to reach the convergence if the data corpus has large amount of missing data.

IV EXPERIMENTAL RESULTS

Performance of the clustering algorithms is measured based on f-measure, recall, precision, purity on three bench mark datasets 20 NG, Reuters-21578 and TDT-2.

Assume Mij – specifies how many number of points present in the class ‘i’ for cluster ‘j’, Mj – Specifies how many number of points present in cluster ‘j’ and Mi - specifies how many number of points present in the class ‘i’.

A. *Precision*

The fraction of appropriate data points among the retrieved data points;

$$\Pr ecision = \frac{M_{ij}}{M_j}$$

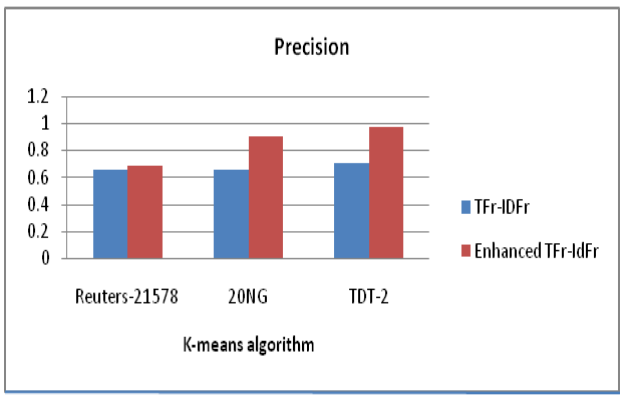

Fig. 1 Comparison of Precision among data sets for K-means clustering algorithm

Figure 1 calculates precision for k-means algorithm with TFr-IdFr and enhanced TFr-IdFr and compare their performance on three datasets Reuters-21578, 20NG and TDT-2.

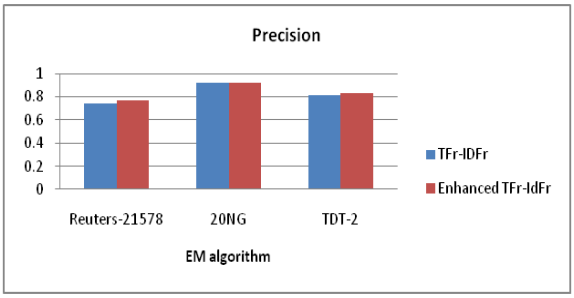

Fig. 2 Comparison of Precision among data sets for EM algorithm

Figure 2 calculates precision for EM algorithm with TFr-IdFr and enhanced TFr-IdFr and compare their performance on three datasets Reuters-21578, 20NG and TDT-2.

B. Recall

The fraction of appropriate data points that have been retrieved over the total amount of appropriate data points

$$\text{Re}call = \frac{M_{ij}}{M_i}$$

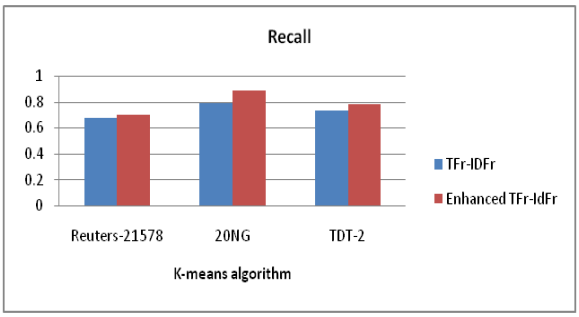

Fig. 3 Comparison of recall among data sets for K-means clustering algorithm

Figure 3 calculates recall for k-means algorithm with TFr-IdFr and enhanced TFr-IdFr and compares their performance on three datasets Reuters-21578, 20NG and TDT-2.

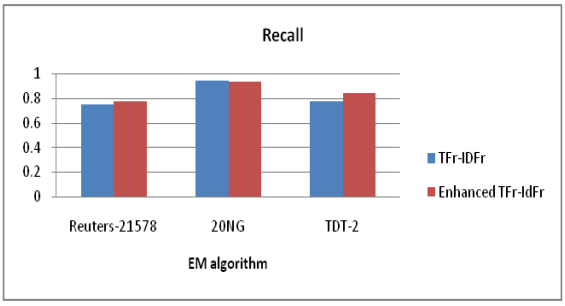

Fig. 4 Comparison of recall among data sets for EM algorithm

Figure 4 calculates recall for EM algorithm with TFr-IdFr and enhanced TFr-IdFr and compares their performance on three datasets Reuters-21578, 20NG and TDT-2.

C. f-measure

It is a measurement technique to measure the accuracy of the clustering algorithm. It uses both precision and recall.

$$F-measure = 2 * \frac{\text{Pr}ecision * \text{Re}call}{\text{Pr}ecision + \text{Re}call}$$

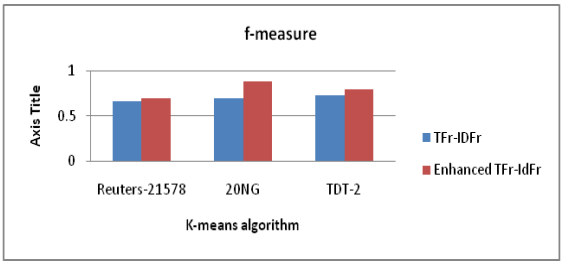

Fig. 5 Comparison of f-measure among data sets for K-means clustering algorithm

Figure 5 calculates f-measure for k-means algorithm with TFr-IdFr and enhanced TFr-IdFr and compares their performance on three datasets Reuters-21578, 20NG and TDT-2.

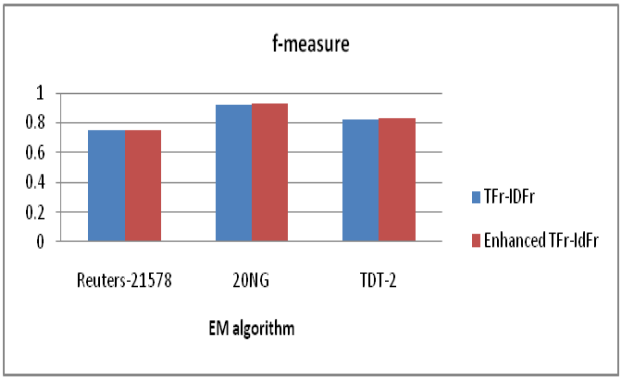

Fig. 6 Comparison of f-measure among data sets for EM algorithm

Figure 6 calculates f-measure for EM algorithm with TFr-IdFr and enhanced TFr-IdFr and compares their performance on three datasets Reuters-21578, 20NG and TDT-2.

D. *Purity*

Let *Mk* specifies how many data points present within the given clusterand let M*lk*specifies how many data points present within the class. Then, the purity ***purity***(*k*) of the cluster is defined as

$$Purity = \frac{1}{M_k} Maximum\ \text{M}_{lk}$$

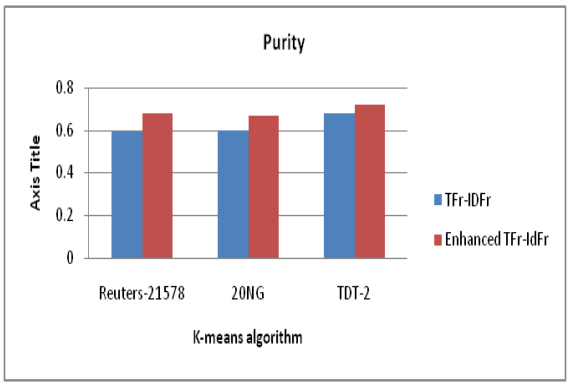

Fig. 7 Comparison of recall among data sets for K-means clustering algorithm

Figure 7 calculates purity for k-means algorithm with TFr-IdFr and enhanced TFr-IdFr and compares their performance on three datasets Reuters-21578, 20NG and TDT-2.

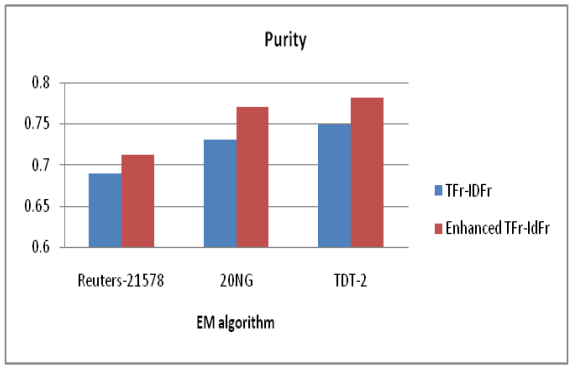

Fig. 8 Comparison of purity among data sets for EM algorithm

Figure 8 calculates purity for EM algorithm with TFr-IdFr and enhanced TFr-IdFr and compares their performance on three datasets Reuters-21578, 20NG and TDT-2.

V CONCLUSION

Text mining is the major area in the present digital world. To extract useful knowledge from the digital data many number of tools are available. An enhanced feature selection is useful for the clustering algorithms to improve the efficiency of clustering. From the observations the clustering algorithms k-means and EM produces better results by using enhanced feature selection compared to traditional feature selection on three bench mark data sets 20 NG ,Reuters-21578 and TDT-2. In future ontology is applied to the clustering algorithms to improve the efficiency of clustering.

REFERENCES

1. Gupta M. and Rajavat A. (2014), "Comparision of Algorithms For Document Clustering", IEEE Sixth International Conference on Computational Intelligence and Communication Networks, (CICN) IEEE computer society, 541-545.
2. Snezhana Salova and bonimir(2017) Incremental clustering algorithm based on phrase-semantic similarity histogram, International Conference on Machine Learning and Cybernetics (ICMLC), Pp. 2088-2093.
3. Cao, T.H., Do, H.T., Hong, D.T. and Quan, T.T. (2008) Fuzzy named entity-based document clustering, Proceedings of IEEE International Conference on Fuzzy Systems, Hong Kong, Pp. 2028-2034.
4. Judith je, Jayakumari j,Distributed Document clustering algorithms : A recent Survey in international journal of enterprise network management 6(3) :207 January (2015)
5. Vikas k vijayan, kr bindu, Latha parameswaran, a comprehensive study of classification algorithms in international conference on Advances in computing ,Communications and informatics, (2017).
6. Poonam Goyal, N.Mehela , Divyansh Bhatia, Topical document clustering : twostage post processing technique,in international journal of Data mining ,Modelling and Management volume no 10, (2018).
7. Boulis C. and Ostendorf M. (2005), "Text classification by augmenting the bag-of-words representation with redundancy compensated bigrams", In Proceedings of the International Workshop on Feature Selection in Data Mining, in conjunction with SIAM SDM-, 9-16.
8. Thangarasu M., Thangamani S. and Manavalan R. (2013) "A Literature Review: Stemming Algorithms for Indian Languages" International Journal of Computer Trends and Technology (IJCTT), 4, 2582-2584.
9. Franca D. and Fabrizio S (2013), "Supervised Term Weighting for Automated Text Categorization", Proceedings of the 2003 ACM symposium on applied computing, ACM New York, NY, USA, 784-788.M.
10. Galavotti, L., Sebastiani, F., & Simi, M. (2006). Feature selection and negative evidence in automated text categorization. Proc. of KDD.
11. Unnati R.Raval ,Chaita jani, implementing and improvisation of k-means clustering algorithms in international journal of computer science and mobile computing ,may (2016).
12. Steffen Barembruch , Anna Scaglione , the expectation and sparse maximization algorithm in Journal Communications (2010).
13. Garima Sehgal, Dr. Kanwal Garg , improved expectation and maximization clustering algorithm in international journal of engineering and computer science,dec-(2017)

AUTHOR PROFILE

VenkataNagarajuthatha Pursuing Ph.D in department CSE, Jawaharlal Nehru Technological University Kakinada.His research interest is Data mining, Machine learning and Deep Learning.

Dr. A. SudhirBabu working as professor in Computer Science and Engineering at PVP Siddhartha institute of Technology. He had Published 28 research articles in reputed International Journals and Conferences. He is acting as a reviewer for 6 International Journals and guiding Ph.D Scholars. Interested in Research areas like Internet of Things, Computer Networks, Cloud Computing and Data Analytics. He published and applied 3 Patents and waiting for results. Dr. A. SudhirBabu is Life member Computer Society of India, ISTE, and Professional member of ACM, IEEE.

Dr. D. Haritha, She is working as Professor in Computer science and Engineering Department at Jawaharlal Nehru Technological University Kakinada. She has 19+ years of experience. She guided 59 M.Tech students and 20 MCA students for their project. Her research interest is on Image Processing, Data Structures, Software Engineering and Networking. She published 15 research papers in international journals. She published 13 research papers in international conferences.