International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT217237

Development of Text Clustering Method with K-Means for Analysis of Text Data

R. J. Wadnare, Dr. S. S. Sherekar, Dr. V. M. Thakare
Sant Gadge Baba Amravati University, Amravati, Maharashtra, India

Article Info
Volume 7, Issue 2
Page Number: 143-151

Publication Issue :
March-April-2021

Article History
Accepted : 01 April 2021
Published : 03 March 2021

ABSTRACT

Clustering is a widely used unsupervised data mining technique. In clustering, the main aim is to put similar data objects in one cluster and dissimilar in another cluster. The k-implies is the most famous clustering algorithm because of its effortlessness. But the performance of the k-means clustering algorithm depends upon the parameter selection. Parameter selection like number of cluster and initial cluster center are key of k-means algorithm. Distance augmentation method, density method quadratic clustering methods are utilized to initial cluster selection. This paper examination five unique methods, for example, improved k-means text clustering algorithm, revisiting k-means, LMMK algorithm, SELF-DATA architecture, Clustering Approach for Relation e.t.c. But these techniques have some limitations. To improve these approach, this paper has proposed the development of text clustering method with k-means for analysis of text data.

Keywords : K-means, tag clustering algorithm, K-means, latent semanticanalysis (LSA), min-max similarity (MMS), Latent Dirichlet Allocation (LDA).

I. INTRODUCTION

Clustering is a broadly utilized unaided information mining procedure. In clustering, the main aim is to put similar data objects in one cluster and dissimilar in another cluster. **The k-implies is the most famous clustering algorithm because of its effortlessness.** But the execution of the k-means clustering algorithm depends upon the variable selection. **Parameter selection like number of cluster and initial cluster center are key of k-means algorithm. Distance augmentation method, density method quadratic clustering method is for the most part utilized to initial cluster selection.** [1] Clustering is widely used for information extraction. In Natural Language Processing (NLP) extracting information from text sources is an important task. Some language technology required text information for better performance. Topic modelling is important for some applications likes (NLP) and information retrieval. It is an unsupervised methodology where a pre-determined number of themes is separated from a specific arrangement of reports on measurable ideas.[2]For convenient use of social media site, the user uses personalize tags and familiar words according to their own understanding. Tag is a keyword that gives more information about the object. Many developers take the advantage of tag information to build personalize tag recommendation systems for users. But there are many problems in tagging system because of its free nature and lack of explicit meaning in the social tag. Different clustering techniques are used in tag development such as K-means and it's improve version, hierarchical clustering, latent semantic analysis (LSA) with clustering. But this technique doesn't use semantic relation between the tags hence less accurate and real clusters are found [3]. The actionable knowledge extraction from text documents is a complex process and required a lot of expertise. In-text mining needs to find previously unknown and implicit data from text documents which include a grouping of data with similar content, topic modelling and detection, clarification model, document summarizations, and document querying. It is a multi-step process that required multiple algorithm implementation and parameters set by the user. It has high computation cost and time consuming because it needs the best joint analysis selection of techniques.[4]Day to day data available on crime is increased. It is not feasible to study that data manually to solve crime related queries. Thusly natural language preparing strategies are most broadly utilized for handling and taking care of such unstructured information for criminal examination. Past strategies utilized in natura language handling are administered procedures and required a ton of human oversight from the criminal business.[5]

This paper zeroed in on five distinct strategies, for example, improved k-means text clustering algorithm, revisiting k-means, LMMK algorithm, SELF-DATA architecture, Clustering Approach for Relation Extraction and propoed improved approach.

II. BACKGROUND

The author have proposed an improved k-means text clustering algorithm by optimizing initial cluster centers. It combines the density method and distance optimization method for determining the initial cluster center. On performing an experiment on different data set the result show that this method improves the stability and accuracy of text clustering.[1]

Some authors have proposed combined method clustering and topic modelling techniques to achieve better Arabic Documents Analysis. It applies the Latent Dirichlet Allocation(LDA) and K-means clustering algorithm to the news documents. It proves by normalizing the weights in vector space dramatically improves cluster quality in text documents. Appling external measures it also compared the combined method with K-means which gives a better result.[2] Authors proposed two algorithms first is MMK which is an improved version of k-means for better centroid selection it uses min-max similarity (MMS). Then it combines MMK with latent semantic analysis (LSA) which gives a new method (LMMK). The new method is more accurate and directive because of considering the semantic relation between tags. The cluster is not directly assigned to the topic hence semantic correction between the topic and cluster can`t find directly. For measuring the performance of MMKS, LSA, and LMMK author build a new tool called (CCR matrix) which shows that LMMKS gives better performance than the other two methods.[3] Author proposed SELF-DATA which is a distributed engine to determine the steps of the data mining process for the collection of documents. It is self -learning engine with the aim to characterize data distribution which simplifies and reduces the collection of text data under process. It works on the Apache Spark framework for parallel and scalable computation of textual datasets. Experimental results on real data sets prove its efficiency to automatically identify a good weight schema and good transformation method with values of specific algorithm parameters.[4]Author proposed an unsupervised approach, with graph-based clustering. It considers the textual corpus of crime against women in India, hierarchical graph-based clustering technique is used to extract relations between name entities in the corpus. The relation between different entities pair is found with the similarity between them based on intermediate context words. The weighted graph has been formed depending on the similarity. To partition graph on edges similarity threshold is set. Entity pairs are group into the cluster by applying an iterative clustering algorithm, cluster characterization is done using the most frequent context word present in it. This method helps to identify the crime pattern, predict future crime, and determine crime prevention strategies.[5]

This paper is focused on five different techniques such as improved k-means text clustering algorithm, revisiting k-means, LMMK algorithm, SELF-DATA architecture, Clustering Approach for Relation Extraction. The paper is as follows. **Section I** Introduction. **Section II** discusses Background. **Section III** discusses previous work. **Section IV** discusses existing methodologies. **Section V** discusses attributes and parameters and how these are affected on clustering techniques, **Section VI** proposed method **Section VII** for outcome result **section VIII** Conclude this paper. Finally **Section IX** gives future scope.

III. PREVIOUS WORK DONE

Caiquan Xiong et al [2016][1] proposed an improved k-means text clustering algorithm by optimizing initial cluster centers. This method has two steps to find clusters. Firstly it finds the initial cluster center and then applies a simple k-means algorithm on selected centers. This method gives a more accurate and stable result.[1]

L Hawarat et al[2017][2] has proposed revisiting k-means and topic modelling , a comparison study to cluster Arabic documents. To cluster Arabic documents it combines clustering algorithm and topic modelling techniques. It measures the correctness of the combined methodology and traditional techniques by applying different external performance measures. It shows that combined methodology outperformed simple clustering techniques.[2]

Jing Yang et al [2017][3] has proposed MMSK-means which is an improved k-means algorithm, and a new tag clustering method called LKMMS. MMSK uses MMS for initial cluster selection and gives stable and accurate clustering results. LMMSK has features of both MMSK- means and LSA because of this it give more directive and accurate clusters. It considers semantic relations between tags. For result comparison, it uses the CCR matrix to compare the result of both proposed methods.[3]

Tania Cerquitelli et al [2017][4] has proposed Data miners' little helper: Data transformation activity cues for cluster analysis on document collections. It builds self-learning data transformation a distributed engine to cluster a collection of text data to form groups of the document represents similar topics that suggest mining steps for analysis of good configuration. [4]

P. DAS et al[2019][5] manifests a graph-based crime analysis scheme that priorities determine the relationship between the entities present in a large corpus. It applies a top-down hierarchical clustering technique in a different domain to better visualization of the crime scene. A complete weighted graph is generated based on the similarity score of the entity. Based on the threshold value the complete graph is partition into sub-graphs. It identifies significant crime patterns in both criminology and the criminal justice industry. [5]

IV. EXISTING METHODOLOGIES

Many clustering plans have been executed in the course of the most recent a very long while. There are various techniques that are executed for various clustering models i.e. improved k-means text clustering algorithm, revisiting k-means, LMMK algorithm, SELF-DATA architecture, Clustering Approach for Relation Extraction.

A] Improved k-means text clustering algorithm: This method uses two techniques to select initial cluster selection. The first is the distance optimization technique and the second is density distance. The data object which is noise data or outlier it removed by calculating the density parameter of each data object in the dataset. Data object with the highest density is selected as the first initial cluster center and the next highest density object which is far from the first center is selected as the second cluster center. After finding the k cluster center traditional k-means clustering technique is [1] Known data collection D={X1 ,X2 ,X3…..Xn}the density parameter of data object xi

$$Dens(x_i) = \sum_{j=1}^{n} u(MeanDist - d(x_i, x_j))$$

The density parameter of data point *x i* is actually the number of data objects in a circle which center is *xi* and radius is Mean-Dis

B] Revisiting k-means:

In this paper revisiting k-means and topic modelling, a comparison study to cluster Arabic documents is proposed. Topic modelling is important for feature reduction and feature selection. To solve the problem of high dimensionality it reduces the Vector space Model (VSM) to simpler by using topic modelling techniques. It also identifies semantic variables in text documents. Secondly, the clustering technique is applied to form clusters. Eternal measures are used to validate the combined proposed method.[2]

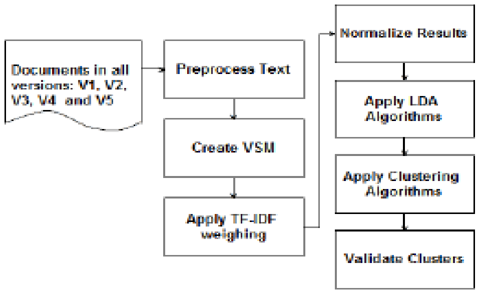
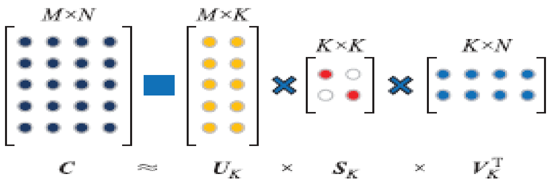


Fig 1. Algorithmfor clustering topics

C] LMMK algorithm:

This paper proposed a new tag clustering method called LMMK. First, it improves the K-means clustering algorithm by choosing the initial centroid with (MIN-MAX Similarity) MMM. The improved method called MMSK gives more stable and accurate results than the traditional k-means algorithm. For more favorable clustering results it builds LMMSK which retains the feature of both LSA with MMSK. To better compare the results between the two proposed methods it finds the CCR matrix of results of both algorithms. applied It is SVD in LSA where K is a number of clusters. Uk is the correction between terms and topics. SK is the correlation degree between terms and documents. Vk is the correlation between documents and topic. [3]



D] SELF-DATA architecture:

The framework of self-data has five building blocks as shown in the figure with the learning phase and prediction phase.

Learning phase: The learning phase responsible for automatically find text clustering process by using previous data characteristics, analysis of several statistical indices to characterize Textual data, store top three results obtains by K-DB, and forecast the feature future analysis through classification algorithm. It includes PASTA, K-DB building prediction model blocks. PASTA is the self-tuning engine used for good weighting schema and analysis of textual collection transformation methods. PASTA includes two steps for proper parameter selection which are document modelling and transformation, self-tuning textual data clustering. K-DB is knowledge base which stores top three results obtained from previous processing document by PASTA. The last advance is building a forecast model which utilizes an order classification on K-DB substance to create a model for great arrangement examination of literary information. A more accurate classification model is built if the larger collection of process collection by self-data is available.

Prediction phase: The contribution to the Prediction stage is a neglected assortment of an archive with a bunch of highlights to portray the printed information appropriation. For the entire clustering process its try to anticipate recommendations. The local and global weight is combining by using a suitable input transformation method to suggest the proper value of the parameter. This method is able to find cohesive, well situated, and correlated groups of documents with similar topics. The entropy of data is defined as$Ej = \sum_{j=1} pij \log (\ pij)$. [4]

E] Clustering Approach for Relation Extraction:

A lot of data is available online on crime. A supervised technique to study crime patterns required more human supervision. It is easy to study a single crime online but difficult to study the pattern of crime in a particular period. To overcome this problem graph-based crime analysis is proposed by the author. Initially, unstructured crime-related data is collected from the newspaper. Name entity recognition module is used to identify different entities such as place, person, and organizations e.t.c. The top-down hierarchical graph-based technique is used to uncover the connection among recognized entities. Entity pairs are classified into three domains namely, PER-PER (person-person), PERLOC (person-location), and ORG-PER (organization-person) for better envision of the crime scene. Intermediate context words and similarity are used as a measure to relation discovery in entity pair in each domain. A weighted undirected complete graph is built from similarity score, where node presented entity –pair and similarity score is represented by the weight of the edge between two nodes. The threshold is calculated using an average of all edge weights. A graph is a partition into two sub-graphs, where one contains edges having a weight greater or equal to a threshold value. Other contains edges with less than a threshold. A threshold is updated for each sub-graph in iteration while partitioning the graph. The procedure is continued till the cluster value is improves, score function cluster validation index is used to measure cluster quality. Finally, a clustering algorithm is applied and the cluster is characterized using the most frequent factors present in them.[5]

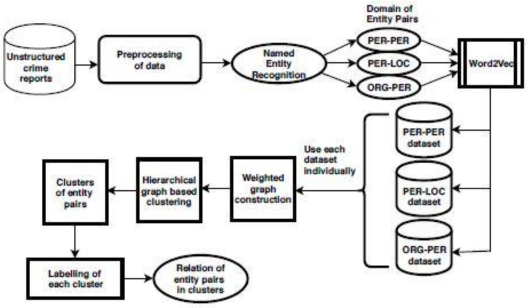

Domain of Entity Pairs
Unstructured crime reports
Preprocessing of data
Named Entity Recognition
PER-PER
PER-LOC
ORG-PER
Word2Vec
PER-PER dataset
PER-LOC dataset
ORG-PER dataset
Use each dataset individually
Weighted graph construction
Hierarchical graph based clustering
Clusters of entity pairs
Labelling of each cluster
Relation of entity pairs in clusters

Fig 2. Flowchart of the proposed methodology.

V. ANALYSIS AND DISCUSSION

For testing the effectiveness and feasibility of the proposed method author select randomly 880 datasets with five different categories from a dataset. It uses Precision and Recalls indices to calculate performance. The trial shows that the accuracy rate and the review pace of the improved K-implies algorithm were fundamentally higher than the customary K-implies algorithm however this technique is tedious.[1] The combined algorithm has achieved an excellent result as compared to the traditional clustering methods. The mean normalization of TF-IDF weight enhances its performance.[2] The experiments are performed in two-stage using MATLAB first stage shows that MMSK-means is faster, stable, and accurate as compared to the original K-means algorithm. The second stage determines that LMMSK performs best over MMKS and LSA –based algorithms using the same dataset.**[3]**By applying the self-Data framework on five real social media data it proved that data spar city is properly described by TTR indices which able to distinguish between sparse datasets and high-density datasets. The datasets having high LSI supports identification more cohesive than PCA.[4] The result computation of this paper is based on three sets of experiments.

Comparison with graph-based clustering algorithms: four graph-based clustering algorithms such as Info map, Louvain, Girvan Newman, Fast greedy are compared with the proposed method. To quantify the viability of the social marking of the clustering structure by graph based techniques Internal just as External group assessment lists are utilized. The outcome shows that interior and outside indices got by the proposed technique are better It gives the finest outcome for the PER-PER domain. For the PER-LOC domain, all other methods have provided better results on the DB index than the proposed method.[5]

Mobility scheme	**Advantages**	**Limitations**
Improved k-means text clustering algorithm	It gives more stable and accurate result. .	It is time consuming technique.
revisiting k-means	It is suitable for high imensionality data sets.	It required large size of text to give acceptable result
LMMK algorithm	It reduces iterations and thus saves time and space cost. It gives more directive and accurate cluster by using semantic relation in tag It guarantees global optimization.	It is time consuming when using singular value decomposition (SVD).
SELF-DATA architecture	It can mine massive data repository with minimal user intervention. It is parameter free technique.	It is complex procedure.
Clustering Approach for Relation Extraction	It gives better semantic analysis of corpus. As it based on unsupervised approach human supervision is less required. It can identify from large corpus and extract	Sometimes label of cluster is not match with all the context word between the entity-pair

	relation among them.	

VI. PROPOSED METHODOLOGY

This paper proposed text clustering method with k-means for analysis of text data. Many times k-means does not give proper result because of wrong value of k. This method pre-processed text data by removing stop-word, punctuation and unwonted words. Words or features are assigned weights according their importance in data sets. Sparse matrix is created by using TFIDF method. Cosine similarity matrix is calculated which gives ideal sparse matrix from TFIDF victimizer which uses non-zero dimension for relation extraction. Latent semantic analysis (LDA) is used for dimension reduction which used important contains in matrix and reduce the size of matrix. The process to normalize the matrix is needs toper form because LSA does not give normalized matrix. MMSK-Means method is applied for global optimization of data. It gives more accurate centroid and number of clusters. And final cluster are found.

Basic steps of algorithm:

Step1. Pre-process the data by removing stop-words, punctuation and unwonted words.

Step2: create the spare matrix by using TFIDF vectorizer with cosine similarity.

Step3: apply Latent semantic analysis (LDA) for dimension reduction.

Step4: apply MMSK-Means method.

Diagrammatic representation of proposed method is shown as follows:

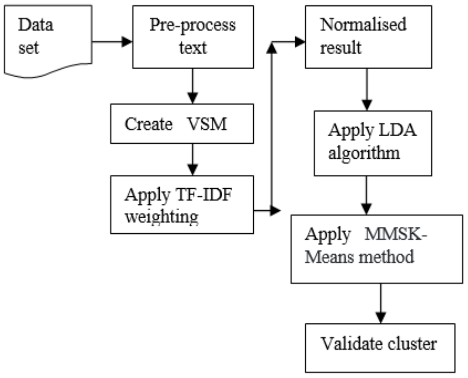


Fig 3. Flowchart of the proposed methodology

VII. RESULT

Test conduct on social book-marking system data of 5 year on MATLAB, compare result of k-means and the proposed method. It gives more accurate result than traditional k-means clustering. The following diagram shows the result of k-means and proposed method.

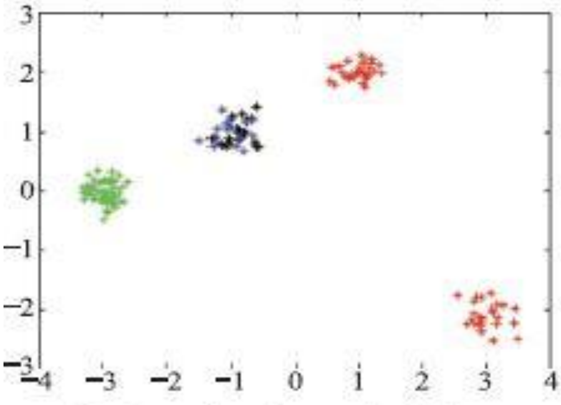


Fig 4. Result of k-means

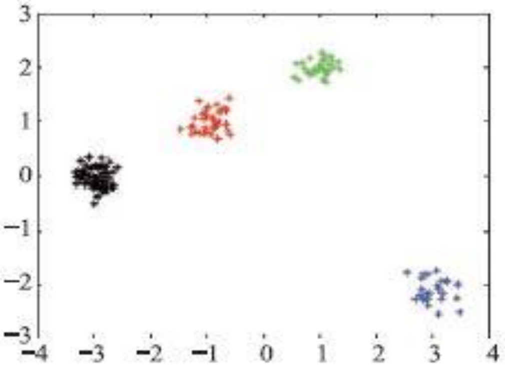


Fig 5. Result of text mining method

VIII. CONCLUSION

Improved k-means text clustering algorithm, revisiting k-means, LMMK algorithm, SELF-DATA architecture, Clustering Approach for Relation methods are study in this paper. But these techniques have some pros and cons. To improve this problem, this paper proposed development of text clustering method with k-means for analysis of text data This method gives more accurate cluster result as it uses mmsk-means over traditional k-means which identify centroid accurately. As it used cosine similarity matrix over correction similarity matrix it gives ideal sparse matrix because it uses non-zero dimension into consideration.

IX. FUTURE SCOPE

Time complexity of the method should be removed.

X. REFERENCES

[1]. Caiquan Xiong Zhen Hua KeLv Xuan Li“An Improved K-means text clustering algorithm By Optimizing initial cluster centers” International Conference on Cloud Computing and Big Data2016

[2]. M.Alhawarat And M. Hegazi“Revisiting K-Means and Topic Modeling, a Comparison Study to Cluster Arabic Documents” IEEE Access2017

[3]. Jing Yang and Jun Wang “Tag clustering algorithm LMMSK: improved K-means algorithm based on latent semantic “Journal of Systems Engineering and ElectronicsApril2017

[4]. Tania Cerquitelli Evelina Di Corso Francesco Ventura Silvia Chiusano“Data miners’ little helper: Data transformation activity cues for cluster analysis on document collections”ACM Reference format June 2017

[5]. P. DAS,A. K. DAS, J. NAYAK, D. PELUSI, W. DING. “A Graph based Clustering Approach for Relation Extraction from Crime Data” IEEE Access.2019

Cite this article as :

R. J. Wadnare, Dr. S. S. Sherekar, Dr. V. M. Thakare, "Development of Text Clustering Method with K-Means for Analysis of Text Data", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 7 Issue 2, pp. 143-151, March-April 2021. Available at doi : https://doi.org/10.32628/CSEIT217237
Journal URL : https://ijsrcseit.com/CSEIT217237

Authors :

Rupali Wadnarehas completed B.E. Degree in computer engineering from RastrsantTukadojimaharaj Nagpur university Maharashtra. She is persuing Master’s Degree in Computer Science and Information Technology from P.G. Department of Computer Science and Engineering, S.G.B.A.U. Amravati. (rupaliwadnare@gmail.com)

Dr. Swati S. Sherekar Presentlyworking as Professor in the P. G.Department of Computer Science andEngg and having 25 years of teachingand research experience. Her area ofresearch is Network security, MobileComputing, Cloud Computing and
working as supervisor for Ph.D.guidance. Completed one MRP of UGC one MRP of AICTE. She has published more than 70 papers in International &amp; National level

Journals and also Invited for Talks in many International / NationalConferences and National levelConferences.(e-mail id:ss_sherekar@rediffmail.com)

Dr. Vilas M. Thakare is Professor and Head in Post Graduate department of Computer Science and Engg, Faculty of Engineering & Technology, SGB Amravati university, Amravati. He is also working as a coordinator on UGC sponsored scheme of e-learning and m-learning specially designed for teaching and research. He is Ph.D. in Computer Science Engineering and completed M.E. in year 1989. He has done his PhD in area of robotics, AI and computer architecture. His area of research is Computer Architectures, AI and IT. He has published more than 150 papers in International & National level Journals and also International Conferences and National level Conferences.
(e-mail id: vilthakare@yahoo.co.in)