Diverse Approaches for Document Clustering in Product Development Analyzer

Mohit Murotiya[1], Madhur Mahajan[2], Ketan Laddha[3], Sourabh Rathi[4], Prof. Shreya Ahire[5]

[1,2,3,4]Student, Department of Computer Engineering, NBN Sinhgad School of Engineering, Ambegaon, Pune, Maharastra, India, 411041
[5]Professor, Dept. of Computer Engineering, NBN Sinhgad School of Engineering, Ambegaon, Pune, Maharastra, India, 411041

---***---

Abstract - *The manual structural organization of documents is expensive in terms of time and efforts. Traversing large number of documents to interpret manually is also challenging issue. Therefore, sophisticated means are needed to cope up with this challenge. Clustering is one of the automated solutions. It is a major tool in many applications of business and data sciences. Document clustering sorts out records into various gatherings called as groups, where the documents in each group share some regular properties as indicated in closeness or similarity measure. This paper proposed method for clustering textual documents using Automatic text classification with TF-IDF, Word embedding algorithm and classifies data using K-means clustering machine learning algorithm.*

Key Words: **Document clustering, Feature Selection, TF-IDF scheme, K-means Clustering, Document organization.**

1. INTRODUCTION

Clustering - an automatic organization of data, which reduces time and complexity to great extent. Clustering is considered as the most important unsupervised machine- learning approach that deals with finding hidden patterns and structures high dimensional unstructured data.

It is impossible manually. Therefore, sophisticated machine learning algorithms are required to perform these tasks for better structuring and getting desired information about data. The machine learning algorithms used for grouping together similar data points (i.e., records, documents) fall into unsupervised learning. The data points are clustered based on certain feature similarity, for example, distance between two coordinate points. It is essential to acquire data into structured format for better understanding and proceeding further for other activities.

Document clustering is a data analysis techniques, which partitions the document into groups of same objects using similarity measure such that similar objects are placed within the same cluster, and dissimilar objects are out of the cluster. The principle of document clustering is to meet human interests in information searching and understanding.

For text documents, the occurrence or count of words, phrases, or other attributes provides a sparse feature representation with interpretable feature labels. In the proposed network, cluster predictions are made using logistic regression models, and feature predictions rely on logistic or multinomial regression models. Optimizing these models leads to a completely self-tuned descriptive clustering approach that automatically selects the number of clusters and the number of feature for each cluster.

2. RELATED WORK

Lot of work has been done in this field because of its extensive usage and applications. In this section, some of the approaches which have been implemented to achieve the same purpose are mentioned.

2.1 Survey on Partition based Clustering Techniques

In this paper, partition based and hierarchical clustering techniques are discussed for clustering of documents with partitions based on clustering algorithms such as K-Means and K-Medoids algorithms, hierarchical clustering, such as Single link analysis and complete link analysis methods. The different similarity measures such as Euclidean distance, Jaccard Coefficient, Cosine similarity, Pearson Correlation and Chi-Squared distance is used for matching the similarity between documents.

K-medoids does not use the mean as the center of the cluster. Instead, it uses medoid. Medoid is the point in the cluster which is most centrally located, and the sum of its distances to other objects or points is the lowest [19]. In each iteration, a randomly picked representative in the present set of medoids is replaced with a randomly picked representative from the collection, if it increases the clustering quality [20][21].

The hierarchical clustering methods form the clusters by recursively dividing the objects in top-down or bottom-up manner. These methods can be again divided according to the manner that the distance function is calculated using SLINK (Single Link Analysis) and CLINK (Complete Link Analysis).

In SLINK analysis, the two clusters distance is considered to be same as the shortest distance between two points such that one point is in one cluster and the other point is in another cluster and in CLINK analysis, the distance between clusters is equal to the distance between the two elements that are at the greatest distance from each other.

2.2 Survey on Clustering using PSO and K-means Algorithm

In this paper, an approach for document clustering using PSO and K-means algorithm is proposed. The algorithm that is proposed includes two major modules, the PSO module and K-means module. At the initial stage, the PSO module is executed for global searching for finding optimal points in the search space. Particles move through the solution space, and after each time step evaluated according to some fitness function. These points are used by the K-means module as initial cluster centroids and then the final optimal clusters of documents are generated. The major steps are Text document collection, document pre-processing, document representation, apply PSO algorithm to initialize centroids for k-means algorithm. The clusters obtained using proposed methods are more compact and isolated from each other than K-means.

2.3 Survey on Document Clustering Challenges

The primary focus of this paper is on the challenges and difficulties that come across when large-scale text documents are clustered To Clustering this excessive amount of data manually undertakes tedious efforts, and it is practically impossible and the machine learning algorithms are not capable of working directly with raw text, therefore the unstructured form of documents has to be transformed into a well-defined structured one. The challenge is to store and manage a large scale of data through a moderate requirement for hardware and software infrastructure.

So, by applying text document clustering techniques over new platforms such as Hadoop and Map-Reduce which can resolve the issues and provides highly accurate and scalable results.

2.4 Survey on Clustering using Semantic Features and Similarity

In this paper, a document clustering method that use the weighted semantic features and cluster similarity is introduced to cluster meaningful topics from document set. The proposed method can improve the quality of document clustering because it can avoid clustering the documents whose similarities with topics are high but are meaningless between cluster and document by using the weighted semantic features. Besides, it uses cluster similarity to remove dissimilarity documents in clusters and avoid the biased inherent semantics of the documents to be reflected in clusters by NMF (non-negative matrix factorization).

2.5 Survey on Text Features Extraction

This paper proposed the TF-IDF statistical model, and use word2vec model and density clustering algorithm to create method to extract text feature, which takes into account both word statistics and semantic features in text. The word2vector model is used to train the word vector in the text, and a new set of text feature vectors suitable for the VSM is generated by clustering those word vector based on the TF-IDF algorithm, which can finally better reflect the text features.

Term frequency–inverse document frequency (TF-IDF), is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. The major steps are Managing Data and Training Word Vector, Exclude low TF-IDF Words, Clustering Similar Words, Calculating the TF-IDF of the word, Constructing Vector Space Model.

2.6 Survey on Multi-Viewpoint Approach for clustering

The architecture of the proposed system is divided into set of modules. Some Pre-processing steps are applied before running clustering algorithms is stop-word removal, stemming, term frequency and tokenization. Then after initialization of k- number of required clusters, cosine similarity is been calculated to determine the objects which has maximum dissimilarity among the documents it belongs.

In this work we have studied the improved incremental k-mean clustering. K-mean clustering algorithm is very popular as well as simple and easy algorithm, but it has some limitation in it also. There has been variety of available algorithm which is ambitious to improve the k-mean algorithm and work around the drawback of the k-mean algorithm.

The k-mean, is based on initial cluster centre selection which has issue for selection of appropriate value of k and cluster centre. The proposed research of improved incremental k-mean can choose correct value of k by selecting high dense objects as cluster centre so that they can provide an efficient and essential clustering for k-mean algorithm. Because of simplicity and ease of understand of k-mean algorithm, makes it choice for many clustering applications. However, k-mean algorithm for document clustering suffers too many problems such as the problem of initializations, dead point problem, and the predetermined number of cluster k. we introduced a novel and suitable method for initializations that aim to find appropriate initial centre for k-mean.

2.7 Survey on Web Document Clustering

If the clusters are predefined than there is no use of finding similarity among documents and they can be directly placed. If the clusters are unknown then clustering is done on the basis of some reference point.

Previous work has utilized cosine similarity and Optimization algorithm based on swarm intelligence for similarity index and cluster optimization. Proposed work enhances the previous work done by the utilization of Artificial Neural Network (ANN) combined with the prior work. The clusters formed would then be validated against classifier.

The proposed model optimizes the clusters for classification and validation process. It provides better results as compared to previous approach where the clusters formed were not validated.

2.8 Survey on K-means Algorithm based on Knowledge Graphs

Proposed an improved K-means algorithm for document clustering which based on the distance to optimize the choice of the initial cluster centroid, which can avoid the drawbacks caused by random selection and adopted the knowledge of graphs to improve traditional k-means text clustering algorithm by optimizing the calculation of text similarity.

2.9 Literature Survey on K- prototype Algorithm

The proposed system consists of preprocessing web document data for removing unwanted data. Next is the feature extraction phase through named entity recognition method and topic modeling approach (LDA). Feature extraction shrinks data dimensionality. K-prototype clustering algorithm approach performs better for clustering as it takes into consideration number of mismatches for categorical data. The execution time and space utilized by K-prototype algorithm is better than Fuzzy clustering algorithms.

The data is clustered using k-prototype clustering which clusters documents, by comparing with features extracted from previous steps. They are calculating execution time for the clustering algorithm by measuring difference between times taken by algorithm before clustering takes place till taken for clustering. Topic modeling gives the topic by which every feature are compared, mismatches are calculated and distances are stored in k-prototype algorithm hence less time is time taken by k-prototype algorithm for execution. The space consumed by K-prototype algorithm is less compared to fuzzy algorithm. Hence the performance of K-prototype clustering algorithm is refined.

2.10 Survey on Extractive text summarization

This paper uses neural networks to perform semantic representations as well as relevance and redundancy checks concurrently that help to improve performance of abstractive text summarizations. Then genetic algorithms or swarm intelligence techniques are implemented to find the best summary of text-documents. In order to increase the scope of summary generation, heterogeneous datasets from multiple domains are used as a source.

Table -1: Analysis of Literature Survey

Survey Papers	Author	Methodology
Performance of Unsupervised Learning Algorithms for Online Document Clustering	Dilip Singh Sisodia, Akanksha Verma	Partition based and hierarchical based techniques for clustering.
An Approach for Document Clustering using PSO and K-means Algorithm	Rashmi Chouhan, Anuradha Purohit	PSO method is used for finding optimal points in search space and these points are considered as initial cluster centroids for K-means method.
Text Document Clustering: Issues and Challenges	Maedeh Afzali, Suresh Kumar	The problems and challenges that come across while clustering a huge amount of text data are discussed.

An improved Document Clustering Approach with Multi-Viewpoint based approach	Anjali gupta, Rahul Dubey	Aim to find appropriate initial centre for k mean.
Web Document Clustering	Vaishali Madaan, Rakesh Kumar	It uses Artificial Neural Network combined with K-means to increase the efficiency of clustering.
Clustering Based on Knowledge Graphs	Xiaoli Wang, Ying Li, Meihong Wang,ZiXiang Yang, Huailin Dong	To improve traditional k-means text clustering algorithm by optimizing the calculation of text similarity.
Concept based document clustering	Sneha Pasarate, Rajashree Shedge	K-prototype clustering algorithm approach performs better for clustering as it takes into consideration number of mismatches for categorical data.
Text Features Extraction	Qing Liu, Jing Wang, NaiYao Wang	Extraction of similar words, and the similarity among words needs to be obtained by the meaning of words in texts.
Extractive Text Summarization Methods	P N Varalakshmi K, Jagadish S Kallimani	This paper describes current methods to perform extractive text summarization where the input would be multi document sets.

3. CONCLUSION

Document clustering is a feasible way of demonstration and it plays an important role in many of the data sciences applications. In this paper we investigated many existing algorithms and different approaches to improve k-means clustering algorithm. This study paper presents various feature extraction techniques adopted in the course of document clustering. Though many algorithms have been proposed for clustering but it is still an open problem and looking at the rate at which the web is growing, for any application using web documents, clustering will become an essential part of the application.

ACKNOWLEDGEMENT

We would also like to show our gratitude to the Dr. Shwetambari Chiwhane, NBN Sinhgad School of Engineering for sharing their pearls of wisdom with us during project work and making of this paper.

REFERENCES

[1] Dilip Singh Sisodia, Akanksha Verma, Performance of Unsupervised Learning Algorithms for Online Document Clustering, ICIRCA.2018.8597378.

[2] Rashmi Chouhan, Anuradha Purohit, An Approach for Document Clustering using PSO and K-means Algorithm, 978-1-5386-0807-4/18/$31.00 ©2018 IEEE

[3] Maedeh Afzali, Suresh Kumar, Text Document Clustering: Issues and Challenges, 978-1-7281-0211-5/19/$31.00 2019 ©IEEE

[4] P. Chahal, M. S. Tomer, and S. Kumar, "Semantic Similarity between Web Documents Using Ontology," Journal of The Institution of Engineers (India): Series B, vol. 99, no. 3, pp. 293-300, 2018.

[5] Anjali gupta, Rahul Dubey, An improved Document Clustering Approach with Multi-Viewpoint based on different similarity measures, 978-1-5386-2842-3/18/©2018 IEEE.

[6] Kudal,Prof. M.M.Naoghare,‖A Review of Modern Document Clustering Techniques‖, International Journal of Science & Research(IJSR), Volume 3 Issue 10, October 2014.

[7] Vaishali Madaan, Rakesh Kumar, An Improved Approach for Web Document Clustering, ICACCCN2018.

[8] A. Bhakkad, S.C. Dharmadhikar, P Kulkarni, and M. Emmanuel, ―EVSM: Novel Text Representation Model to Capture Context-Based Closeness between Two Text Documents‖, IEEE International Conference on Intelligent Systems and Control (ISCO), Coimbatore, India, pp. 345-348, 2013.

[9] Xiaoli Wang, Ying Li, Meihong Wang, ZiXiang Yang, Huailin Dong, An Improved K-means Algorithm for Document Clustering Based on Knowledge Graphs, 978-1-5386-7604-2/18/$31.00 ©2018 IEEE

[10] Macqueen J. Some Methods for Classification and Analysis of MultiVariate Observations. Proc. of, Berkeley Symposium on Mathematical Statistics and Probability. 1967:281-297.

[11] Shwetambari Kharabe, C. Nalini," Robust ROI Localization Based Finger Vein Authentication Using Adaptive Thresholding Extraction with Deep Learning Technique", Jour of Adv Research in Dynamical & Control Systems, Vol. 10, 07-Special Issue, 2018.

[12] Sneha Pasarate, Rajashree Shedge, Concept based document clustering using K prototype Algorithm, 978-1-5386-0796-1/18/$31.00 ©2018 IEEE

[13] Chun-Ling Chen , Frank S.C. Tseng , Tyne Liang , ―An integration of WordNet and fuzzy association rule mining for multi-label document clustering‖, Data and Knowledge Engineering 69, pp. 1208-1226, 2010.

[14] Wajiha Arif and Naeem Ahmed Mahoto, Document Clustering – A Feasible Demonstration with K-means Algorithm, 978-1-5386-9509-8/19/$31.00 ©2019 IEEE

[15] Jun, Sunghae, Sang-Sung Park, and Dong-Sik Jang. "Document clustering method using dimension reduction and support vector clustering to overcome sparseness." Expert Systems with Applications 41.7 (2014): 3204-3212

[16] Wang, Ye, et al. "Semi-supervised collective matrix factorization for topic detection and document clustering." 2017 IEEE Second International Conference on Data Science in Cyberspace (DSC).IEEE, 2017

[17] P N Varalakshmi K, Jagadish S Kallimani, Survey on Extractive Text Summarization Methods with Multi-Document Datasets, 978-1-5386-5314-2/18/$31.00 ©2018 IEEE

[18]S. Ma, Z.-H. Deng and Y. Yang, "An Unsupervised MultiDocument Summarization Framework Based on Neural Document Model," in Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, Osaka, Japan, 2016.