International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

Document Clustering using Improved K-means Algorithm

Anjali Vashist[1], Rajender Nath[2]

[1]Department of Computer Science and Applications, Kurukshetra University, Haryana, India

[2]Professor, Dept .of Computer Science and Applications, KurukshetraUniversity,Haryana, India

Abstract: ***Clustering is an efficient technique that organizes a large quantity of unordered text documents into a small number of significant and coherent clusters, thereby providing a basis for intuitive and informative navigation and browsing mechanisms. It is studied by the researchers at broad level because of its broad application in several areas such as web mining, search engines, and information extraction. It clusters the documents based on various similarity measures. The existing K-means (document clustering algorithm) was based on random center generation and every time the clusters generated was different In this paper, an Improved Document Clustering algorithm is given which generates number of clusters for any text documents based on fixed center generation, collect only exclusive words from different documents in dataset and uses cosine similarity measures to place similar documents in proper clusters. Experimental results showed that accuracy of proposed algorithm is high compare to existing algorithm in terms of F-Measure, Recall, Precision and time complexity.***

Keywords: Document Clustering, Cosine Similarity, Term Finder, Tf-Idf, Threshold

1. Introduction

Numbers of users on the Internet are increasing day by day and information on the Web is growing exponentially. Extracting useful information from the Web is becoming increasingly difficult. To address this issue, many data mining techniques have been come. Data mining is the process of extracting the implicit, previously unknown and useful information from data. Document clustering, the sub-problem of data mining is the process of organizing documents into clusters and the documents in each cluster share some common properties according to similarity measure used. The document clustering algorithms play an important role in helping users to speedily navigate, sum up and organize the information.

Document clustering can be used for classification of different documents, detecting content duplicity, recommending Web pages to users, optimizing searches etc. Vector space model is used by our document clustering technique. Preprocessing is done to convert the words to their base form,to remove stop words, duplicate words before applying vector space model to the text documents. The distances between documents are measured using similarity measures like Cosine, Euclidian Jaccard etc. Then the clustering algorithms are applied till required number of clusters is formed. There are two common clustering algorithms. Partitioning algorithms in which clusters are computed directly. Clustering is done by iteratively swapping objects or groups of objects between the clusters and the hierarchical based algorithms in which a hierarchy of clusters is build. Every cluster is subdivided into child clusters, which form a partition of their parent cluster. Different clustering algorithms [1] produce different results with different features. Hierarchical algorithms are slower than partitioning algorithms but they give better accuracy. Experimental results showed that theproposed algorithm takes less time for clustering compare to existing K-means algorithm and the F-measure score that is F-1 score of clustering, Recall, Precision is also very high than existing.

The remainder of this paper is organized as follows. Section II presents related work. Section III describes research methodology Section IV explains details of proposed system. Section V analyzes the Experimental results. Conclusion and Future Work is given in Section VI.

2. Related Work

K-means clustering comes under partitioning clustering algorithm. It partitions given data into K clusters. Several other clustering algorithms are proposed for dealing with document clustering. Novel algorithm [2] for automatic clustering suggested how clustering is done automatically, improved partitioning K-means algorithm[3] presented newmethod for initializing centroids. Ontology based k-means algorithm [4] presented how ontological domains are used in clustering documents. Improved document clustering algorithm using K-means [5] presented solution of over clustering by partitioning of documents using divide and conquer approach.

Above discussed methods [2],[3],[4],[5] used 20newsgroup, Rueters-21578 and Real time data sets. All Algorithms used cosine similarity measure for finding similarity. Initial k-value is specified in every algorithm except first and last one because first and last algorithm is automatic. They considered documents categories for automatic clustering. Zero clustering is the major advantage of first algorithm. i.e. documents with zero value in similarity matrix also get cluster. Remove over clustering is the major advantage of fourth algorithm. All Algorithms are adaptable to dynamic data except second, because second algorithm calculates average document similarity matrix. After analysis we come to know that these algorithms have limitations, i.e. random center generation, not consider the semantic analysis of documents. The improved k means algorithm has a major

limitation it takes nonexclusive words also and do not match the words by semantic basis. To overcome these limitations a new algorithm is developed.

3. Research Methodology

Getting relevant data from a collection of documents is not merely a single operation, but is more a process in multiple stages. These stages include more traditional information retrieval operations such as crawling, indexing, weighting, filtering etc. Some of these other processes are central to the quality and performance of most clustering algorithms, and it is thus necessary to consider these stages together with a given clustering algorithm to harness its true potential. I will give a brief overview of the clustering process, before we begin our literature study and analysis. Figure 1 represents the various major steps involved in document clustering process. Offline document clustering approach can be divided into four stages given below:

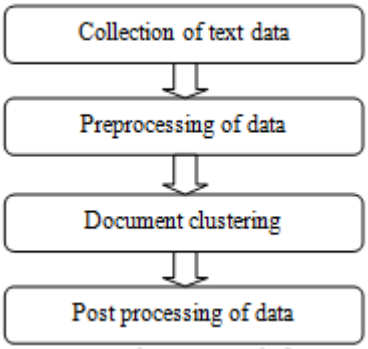

Figure 1: The Stages of the Process of Document Clustering

Collection of text data includes the processes like crawling, indexing, filtering etc. which are used to collect the documents that need to be clustered, index them to store and retrieve in a better way, and filter them to remove the extra data, for example stop words.

Pre-Processing of data is done to represent the data in a form that can be used for clustering. There are many ways of representing the documents like, Vector-Model, graphical model, etc. Many measures are also used for weighing the documents and their similarities.

Document clustering is the main focus of this thesis work and will be discussed further.

Post-Processing of data includes the major applications in which the document clustering isused, for example, the recommendation application which uses the results of clustering for recommending news articles to the users.

4. System Architecture of Proposed System

System architecture of the proposed document clustering methodology is as given in following Figure 2.

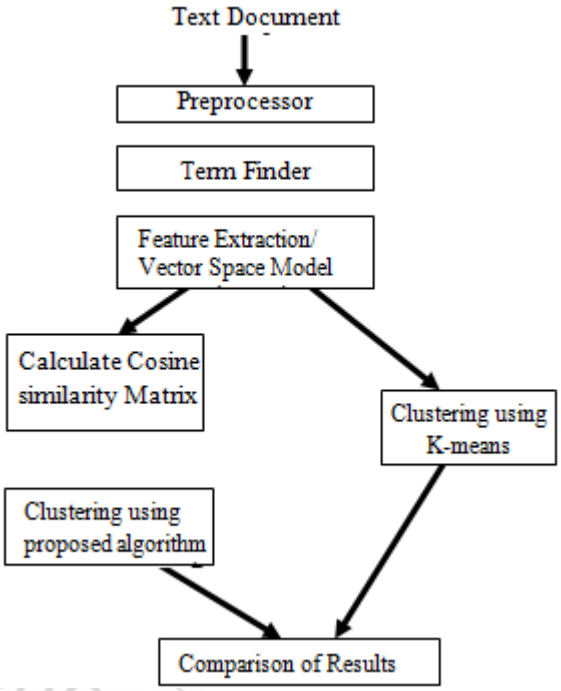

Figure 2: System Architecture of Proposed System

The proposed system follows the sequence of steps as shown in figure 2. System takes text document dataset as input. Pre-processing is done on all documents to remove some non-informative words e.g. “in,and, the”. Term finding removes all the nonexclusive words from text documents and generates the exclusive words. Feature vector is extracted from the dataset by taking keywords with maximum frequency. Semantic analysis is performed to retrieve the keywords which are semantically same by using Word Net library. Vector space model (VSM)[6] computes the term frequency and weight of each word in the documents. Final vector space model is a numerical matrix representation of text data. Cosine Similarity matrix is calculated. These matrices are then used as input to K-means and proposed algorithm and clustering is done. Finally results are compared for different parameters like F-measure, time complexity.

Documents are pre-processed in order to reduce the dimensions of the document vector space. To achieve this main steps involved are stemming, term filtering, stop word removal, tokenization and document representation. There are many morphological variants of words have similar meaning can be considered equivalent. Filtering process removes the special characters and punctuation from the text document which are not hold any discriminative meaning in vector space model. The removal of stop words is the most common term filtering technique used. There are standard stop word lists available but in most of the applications these are modified depending on the quality of the dataset. Some other term filtering methods are:

- Removal of terms with low document frequencies. This is done to improve the speed and memory consumption of the application.
- Numbers do not play much importance in the similarities of the documents except dates and postal codes. Thus these can also be removed.

- Removal of general words, adverbs and non-noun verbs which generally do not make the context of the user query.
- Removal of this that possess less than 3 characters.

A stop word is produced as a term which is not thought to represent any meaning as a dimension in the vector space. Stop words are the most common words (e.g., "and", "or", "in") in a language, but they do not convey any significant information so they are stripped from the document set. This step splits sentences into individual words. The various clustering algorithms use the vector space model to represent each document. In this model, each document d is considered to be a vector in the term space. In its simplest form, each document is represented by the term-frequency (TF) vector.

In this proposed approach of document clustering will produce high-quality document clusters, a process of clustering orgrouping of textual documents based on rules. This proposed method focuses on frequent term similarity measures, andsequence frequent terms. Thismethod introduces a new similarity measure based on frequent terms. Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. This is the most commonly used method to compute cosine similarity between the documents. To compute cosine similarity we use term frequency vector of the documents.

Two vectors with the same orientation i.e. angle 0^0 have a cosine similarity of 1, two vectors at 90^0 have a similarity of 0, and two vectors diametrically opposed have a similarity of -1, independent of their magnitude. Cosine similarity [7] is particularly used in positive space, where the outcome is neatly bounded in [0, 1].

Similarity = $\cos(\theta) = \frac{A.B}{||A||||B||} = \sum_{i=1}^{n} \frac{A_i * B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$

Length normalization of each document vector by the Euclidean length of the vector turns the vector into unit vectors. In doing so, all information on the length of the original document is eliminated. Document length normalization is a way of penalizing the term weights for a document in accordance with its length. Some of the normalization techniques used in information retrieval systems are cosine normalization, Byte Length normalization etc. Cosine normalization is the most commonly used technique in the vector space model. In vector space model, a document is characterised by a vector where the value of each dimension corresponds to the number of times that term appears in the document. Cosine similarity between two documents gives a useful measure of how similar two documents are likely to be in terms of their subject matter. To calculate cosine normalization L2 norm is calculated first.

L_2 norm = $\|\vec{X}\| = \sqrt{\sum_i X_i^2}$

Dividing vector by its L2 norm makes it a unit (length) vector. Suppose$\vec{d_1}, \vec{d_2}$length normalized vectors then cosine similarity between$\vec{d_1}$, $\vec{d_2}$ is simply the dot product of $\vec{d_1}$, and $\vec{d_2}$

$Cos(\vec{d_1}, \vec{d_2}) = \vec{d_1}.\vec{d_2} = \sum_{i=1}^{|V|} d_{1(i)} * d_{2(i)}$

Proposed methodology also uses term frequency vector of each document for the purpose of calculating similarity matrices for each proposed algorithms. In the proposed method we normalize the match vector space by dividing the vector by size of the smaller, larger or average of the two vectors. Match vector between two term frequency vectors is a vector which counts the number of times corresponding term matches.

A. Proposed Algorithm of System

Input: Dataset set D = {d1, d2 . .. dn}
Output: Set of Cluster Numbers C along with document numbers m associated.

1. U= {DI I i € N}
2. Now apply improved K- means algorithm on every partition iteratively till we get the same clusters.

2.1 Input the cluster from user i.e. K (no of clusters).
2.2 Sort the Vector Space Model (VSM) and generate the K parts.
2.3 Take mean of every column (i.e. mean of every part)
2.4 The mean calculated is center of prediction.

3. Calculate the similarity of the documents using cosine similarity measure.
4. Assign the nearest (similar) document to the new clusters.
5. If the clusters are not matched then go to step 3.
6. If clusters are matched then stop.

5. Results and Discussion

A. Datasets

The English data set is the popular collection called 20 Newsgroup. The documents are almost evenly distributed over the different newsgroups. There is a modified version where duplicated and cross-post have been removed, which we have used throughout our work. The headers only contain from and subject fields and this reduced data set contain 20000 documents. In modern search application it is common to generate keywords from the indexed data and display these when presenting the query results. The keywords are a set terms or noun phrases that represent a document. The idea behind this experiment is that we shall preprocess our data and extract keywords from our documents and allow for these keywords to represent the documents. Its details are as follows:

- Number of unique documents = 18,828
- Number of categories = 20

B. Results of System

Existing algorithm takes more time than proposed algorithm. Existing algorithm takes more time because it also takes non exclusive words for clustering of text documents and do not match the documents semantically. Proposed algorithm takes less time because it takes only exclusive words from dataset and matches the words semantically.P (Precision) is the ratio of the number of relevant documents to the total number of documents retrieved for a query. R (Recall) is the ratio of the number of relevant documents retrieved for a query to the total number of relevant documents in the entire collection. A measure that combines precision and recall is the harmonic mean of precision and recall.

Table 5.1: Quality Comparison of Existing and Proposed System

Results of Existing System		
	CLASS 1	CLASS 2
Cluster 0	54	63
Cluster 1	46	37
Results of Proposed System		
	CLASS 1	CLASS 2
Cluster 0	53	0
Cluster 1	47	100

Table 5.2: Performance Comparison of Existing and Proposed System

Results of Existing System			
	Precision	Recall	F-Measure
Class 1	0.538462	0.63	0.580645
Class 2	0.554217	0.46	0.504732
Results of Proposed System			
	Precision	Recall	F-Measure
Class 1	1	0.53	0.69281
Class 2	0.6802	1	0.8097

Table 5.3: Time Comparison of Existing and Proposed System

	Existing	Proposed
Time (ms)	109	62

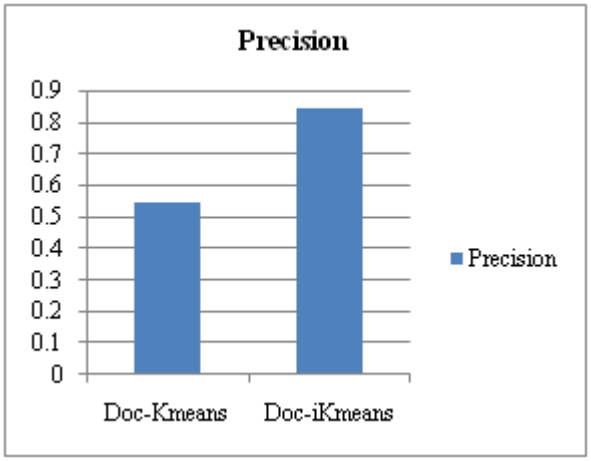

Figure 3: Precision Comparison between K-Means and Improved K-Means

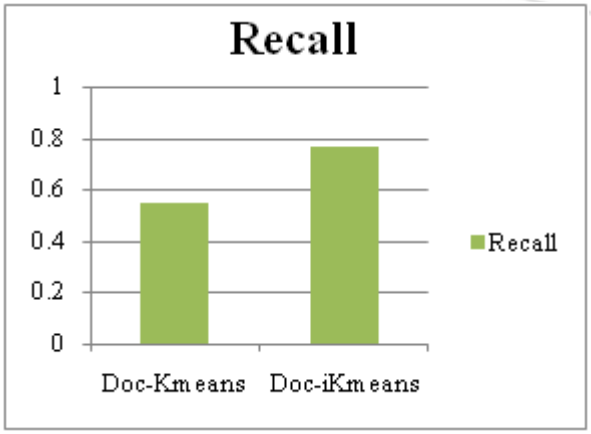

Figure 4: Recall Comparison between Existing K-Means and Improved K-Means

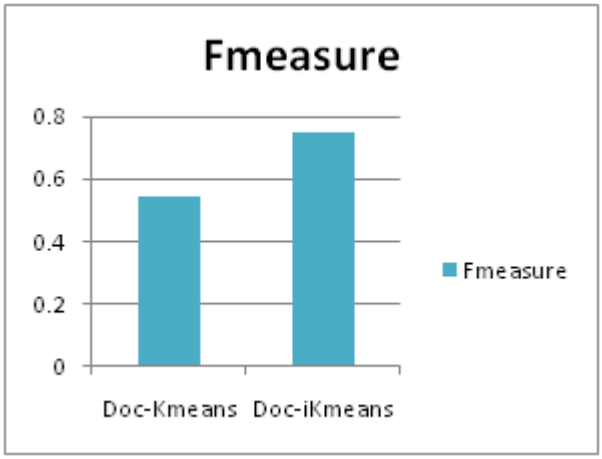

Figure 5: Fmeasure Comparison between Existing K-Means and Improved K-Means

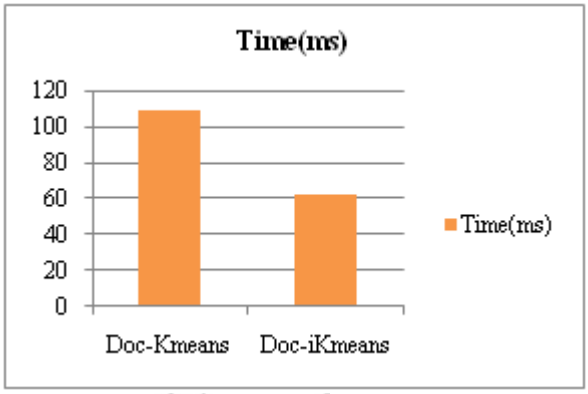

Figure 6: Time comparison between existing k-means and improved k-means

6. Conclusion & Future Scope

A. Conclusions

Document clustering plays an important role in selecting required documents among the thousands of documents.The proposed clustering process improves the quality of data clusters that will provide users with knowledge about the content of a document as well as produce a close textual data clusters to natural classes. Additionally frequent term based clustering approach improves the performance of the system and of the clustering process. In this thesis some existing algorithms is investigated and an improved one is proposed. Existing algorithm takes more time than proposed algorithm. Existing algorithm takes more time than proposed algorithm. Existing algorithm takes more time because it also takes nonexclusive words for clustering of text documents and do not match the documents semantically. Proposed algorithm takes less time because it takes only exclusive words from dataset and matches the words semantically.

B. Future Scope

If keyword searching can be incorporated with the document clustering then the document grouping and retrieval will be more efficient and less time consuming. Comparing the performance of similarity measures using various other clustering algorithms can also be done as a future work. The algorithm proposed in this thesis is just improved in one or two steps and there may be many possible improvements that can be implemented.

References

[1] A.K Jain, A. Topchy, M.H.C Law, J.M Buhman," Landscape of Clustering Algorithms",Seventeen International Conference on Pattern Recognition, Pages 260 - 263 Vol.1, 2004IEEE
[2] Zonghu Wang, Zhijing Liu, Donghui Chen, Kai Tang,"A NewPartitioning Based Algorithm For Document Clustering",EighthInternational Conference on Fuzzy Systems and KnowledgeDiscovery,pages 1741 - 1745 IEEE,20 I I.
[3] S.C. Punitha, R. Jayasree andDr. M. Punithavalli, "Partition DocumentClustering using Ontology Approach", Multimedia and Expo, 2013International Conference on Computer Communication and Informatics(ICCCI -2013), Jan. 04 06,pages 1-5, 2013.
[4] Lus Filipe da Cruz Nassif and Eduardo Raul Hruschka, "DocumentClustering for Forensic Analysis: An Approach for Improving ComputerInspection," IEEE transactions on information forensics and security,Vol. 8, NO. I ,pages 46 - 54 Jan 2013.
[5] Pramod Bide, "Improved Document Clustering using K-means Algorithm", conference on Electrical, Computer and Communication technology, pp. 1-5, 5-7 March 2015 IEEE.
[6] Dharmendra Sharma, Suresh Jain," Context-based weighting for vector space model to evaluate the relation between concept and context in information storage and retrieval system", International conference on computer communication and control, Pages-1-5, 2015 IEEE.
[7] O.Egecioglu, H. Ferhatosmanoglu, U.Ogras," Dimensionality reduction and similarity computation by inner-product approximations", Transactions on Knowledge and Data Engineering, Vol-16, Issue-6, Pages-1041-4347, 2004 IEEE.