International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 4 Issue: 1 41- 44

Study of Document Clustering Methods

Mr.Mayur Bhosale
Dept.Computer Engg,
RAIT, Navi Mumbai, India
mayurbhosale511@gmail.com

Prof.Tushar Ghorpade
Dept.Computer Engg,
RAIT, Navi Mumbai, India
tushar.ghorpade@gmail.com

Prof.Rajashree Shedge
Dept.Computer Engg,
RAIT,Navi Mumbai, India
rajashree.shedge@rait.ac.in

Abstract: Cluster analysis or clustering is the task of grouping a set of object in such a way that object in Same group are more similar to each other than to those in other group. All the data available on the Internet or in forensics is basically stored in different formats of documents. This data (i.e. documents) are huge. So, in order to maintain all this document in a proper clusters document clustering is needed. K-means is one algorithm used in document clustering. For finding similarity used cosine similarity and Jaccard coefficient .In this report from comparative analysis we come to the conclusion that using K-Means algorithm with jaccard coefficient best result outcome in Document clustering process.

Keywords: *Clustering-mean algorithm, Cosine similarity, jaccard coefficient, Improve K-mean algorithm*

*****

I. INTRODUCTION

Document Clustering is an important issue in text mining. Clustering has been widely applicable in different areas of science, technology, social science, biology, economics, medicine and stock market. Clustering problem appears in other different field like pattern recognition, statistical data analysis, bioinformatics, etc. Cluster is a group of similar objects. In other aspects good document is one where intra cluster distance is minimum and maximum distance between inter cluster. The main thing in clustering is no need of human expertise methods which different ate it from the classification.

Clustering method also depends on the document representation technique. We have met with the next two document clustering method. First is Vector space Model. Second is Matrix Representation. VSM is used for to represent documents and web pages. VSM displays the collection of documents in matrix where row represents the number of documents and column number represents overall document term. The Matrix Representation Technique represents each document as matrix Mi and the number of rows showing segment within documents. The main aim of this two technique is to find the matching keywords in the group of document [1]. Then the clustering algorithms are applied till required number of clusters is formed.

Basically there are two types of clustering – hierarchical clustering and partitioning clustering [2] .In partitioned clustering data is divided into different groups. In hierarchical clustering data is divided into two sub group at every rotation means is one of the partitioned based clustering method. In which the different cluster are found based on centroids.data items are moving their position by nearest centroid.Nikhil Chaturvedi developed new K-means algorithm. In which the centroid find systematically .Due to that the accuracy and time improved .The high complexity and low accuracy are still issues and challenges in the clustering. This motivates the study of Document Clustering. The remainder of this paper is organized as follows. Section II presents related work and literature survey. Section III document clustering techniques in detail. Section IV Comparison and analysis .Section V concludes the report.

II. RELATED WORK

Most of the clustering methods depend on various preprocessing techniques to achieve optimal quality and performance. We discuss here some of the common preprocessing methods [4].

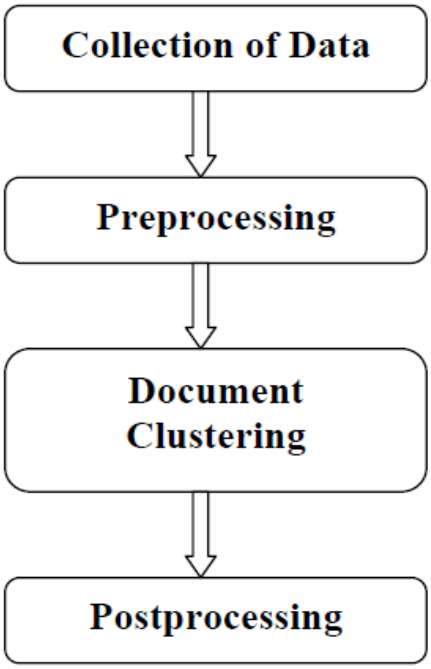

Figure 1.Stages of process of clustering [4]

In first stage process doing like indexing, crawling, etc.Also is used to collect the document which is to be clustered. In next stage find the similarity based on weighting the document. Also is used to representing the data. In next stage main part is clustering, will discussed in details. Last stage is a application for that clustering is used. In addition to the various clustering techniques discusses about various document-representing techniques in graph. In particular Vector Space Model and

Matrix Representation, in improved K-Means algorithm use systematically finds initial centroid which reduces the number of data base scan and it useful for large amount of data base scan .This method reduced time for execution[2].Similarity measurement method between words by deploying Jaccard Coefficient and TFIDF. In tf-idf weight is composed by two terms. In fist computes the normalized Term Frequency(TF)by calculating the number of times a word appears in a document, divided by the total number of words in that document. In second term Inverse Document Frequency (IDF), computed as the logarithm of the number of the documents in the corpus divided by the number of documents where the specific term appears[10].

III. DOCUMENT CLUSTERING TECHNIQUES

There are many techniques used for clustering the documents.In this Chapter various document clustering techniques are explained.

1.1 K-means clustering algorithm

As per the K-menas clustering algorithm first choose cluster center. Secondly calculate the distance from data value and cluster centers and assign it to the nearest cluster. Lastly update the cluster data values. Do the same task until the new cluster do not formed [2]. Figure 2 shows how to process of the basic k means clustering algorithm.

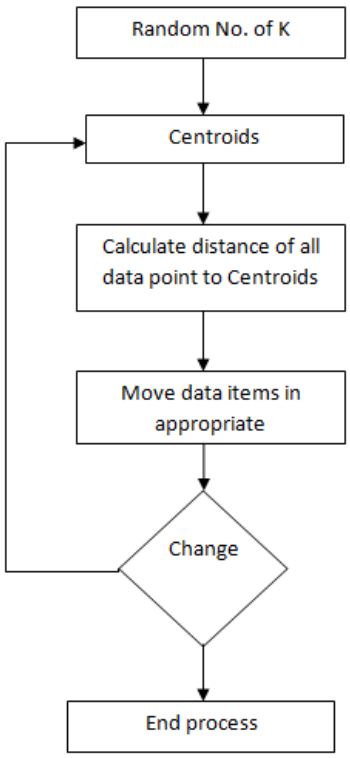

Figure 2. K-means algorithm process[2]

Algorithm:

1) Select K points as temporary centroids
2) Repeat
3) Form K cluster by assigning points to its closest centroid (having smallest distance)
4) Compute the centroid vector C of each cluster.
C=1/N here N is all data. C stands for centroid; N is the number of document in particular cluster.
5) Until centroids do not change.

1.2 Jaccard coefficient for keywords similarity

Thus algorithm is for similarity measurement method between words and index terms. Jaccard methods explain coefficient value in between 0 and 1.0 means there is no similarity and 1 means similarity [5]. Some steps to calculate the similarity as follows

- The data relationship between the information

1) The information prepared as words which were grammatically correct.
2) The information that was not grammatically correct was tested in three groups (the misspelled words, crashed words, and over-typed words).

- A measure of similarity of the search words

1) Measuring the Jaccard similarity coefficient between two data sets is the result of division between the number of features that are common to all divided by the number of properties.
.
2) To find the matching words.

1.3 Clustering based on cosine similarity measure

procedure INTIALIZATION Select k seeds s1,.sk
randomly Cluster[di] ← p = argmaxr{srtdi};
i = 1…..n
Dr didi € sr; nr ← | sr |; r = 1,2....n end procedure
procedure REFINEMENT repeat
{ v[1:n] } random permutaion of {1,2.....n }
for j ←1:n do i ← v[j]
p ← cluster[di]
d ̀Ip ← I np-1, Dp-di-I(np,Dp)
q ← argmaxr, r≠p {I nr+1, Dr+di - I nr, Dr}
d ̀Iq ← I nq+1, Dq+di-I(nq,Dq)
if d ̀Ip+ d ̀Iq ≠0 then
Move di to cluster q : cluster di ← q
Update Dp,np,Dq,nq
End if
end for until No move for all n documents
end procedure

1) Initializing the weights parameters.
2) Using the EM algorithm to estimate their means and covariance.
3) Grouping the data to classes by the value of probability density to each class and calculating the weight of each class.
4) Repeat the first step until the cluster number reaches the desired number or the largest OLR is smaller than the predefined threshold value. Go to step 3 and output the result. A distinctive element in this algorithm is to use the overlap rate to measure similarity between clusters[3].

1.4 Improved K-means

In improved k-means algorithm there are two phases which are following

The Initial Centroids

1) Set p = 1;
2) Measure the distance between each data and all other data in the set D;
3) Find the closest pair of data from the set D and form a data set Ap (1 ← p ← k) which contains these two data, Delete these two data from the set D;
4) Find the data in D that is closest to the data set Ap, Add it to Ap and delete it from D;
5) Repeat step 4 until the number of data in Ap reaches all data in D;
6) If p¡k, then p = p+1, find another pair of data from D between which the distance is the small form another data set Ap and delete them from D, Go to step 4;
7) for each data-point set Ap (1 ← p ← k) find the mean of data in Ap.

Data to the clusters

Input: D-n data set. C-Different centriods.
Output: A -More than one cluster.
1) Calculate the distance from the data to the all centroids.
2) find the nearest centroid to each data.
3) Set Cluster CL[i] = j; // j: CL of the closest cluster
4) Set Shorter Dist[i] = d (di, cj);
5) For each cluster j (1 ← j ← k), recalculate the centroids;
6) Repeat
7) For each data di,
Compute the distance from the centroids of the closest cluster;
If distance is less than or equal to the present closest distance, the data-point stays in cluster;
Else
For every centroid compute the distance
End for;
Data di assign to the cluster with the closest centroid cj
Set Cluster CL[i] =j;
Set Shorter Dist[i] = d (di, cj);
End for;
8) For each cluster j (1 ← j ← k), recalculate the centroids;
until the criteria is met.

In the first phase determine initial centroids systematically. In next phase use new approach to improve accuracy .Also finding the clusters based on the calculating distance from centroids to each data[2].

IV. COMPARITIVE ANALYSIS

We have selected four existing methods of document clustering for comparison. Basic facts governing Document Clustering

Measurement of similarity: First and the foremost things to be considered before clustering is the distance measure.The measure reflects how close the target objects are[6]. The nature of similarity measure plays a very important role in the success or failure of a clustering method. some of the similarity measures explained briefly below based on single view point and multiviewpoint.

1) Cosine Similarity

It quantifies correlation between vectors ta and tb as cosine of the angle between them in m-dimensional space. Bounded between [0,1] and independent of document length[8]. where ta and tb are m-dimensional vectors over the term set T = t1, ..., tm.

2) Jaccard Correlation The intuition for this measure is that it measures similarity as the intersection divided by the union of the objects. Formally, it is given by: The Jaccard coefficient is a similarity measure and ranges between 0 and 1. It is 1 when ta = tb and 0 when ta and tb are disjoint, where 1 means the two objects are the same and 0 means they are completely different. The corresponding distance measure is DJ = 1 SIMJ we are
going to compare results of fourth algorithm [9].

Table 1: Comparative analysis of clustering methods

Parameters	Basic K-means algorithm	Cosine Similarity	Jaccard Coefficient	Improve K-Means algorithm
Equality Measure	Euclidean	Distance	Cosine similarity	Jaccard Coefficient
Purity	Less	Less	More	Less
Execution time	Less	More	More	Less
Starting predicates	Required	Required	Required	Required
Possibilities of Clustering	K-clusters	K-clusters	K-clusters	K-clusters
Adaptability	No	Yes	Difficult	Yes
Correctness	Less	Less	Better	Better

Above discussed methods in table I, In document clustering cosine similarity and jaccard coefficient measure for finding similarity. In this Jaccard Coefficient is better due to purity. Starting predicates is required in all the algorithms because algorithm are not automatic [7]. Correctness and execution time of improved k-means algorithm is better than basic K-means algorithm because in improved K-means algorithm use the systematically calculation of centroid instead of random assigned due to which accuracy and time improved. So, from comparative analysis we come to the conclusion that using K-Means algorithm with jaccard coefficient best result outcome in Document clustering process.

V. CONCLUSION

Classical document clustering algorithms use the selecting the initial centroids approach. It will be work on short size of data and to finding centroids problem. In improved K-Means algorithm use systematically finds initial centroid which reduces the number of data base scan and it useful for large amount of data base scan .This method reduced time for execution. Efficiency of the information retrieval is enhanced

by using the Cosine similarity and Jaccard coefficient similarity measures.Using K-Means algorithm with cosine similarity or Jaccard coefficient as input fetches the best result outcome in the Document clustering process. With comparative studies Jaccard coefficient is better due to purity.

REFERENCES

[1] Chandan Jadon and Ajay Khunteta, A New Approach of Document Clustering. International Journal of Advanced Research in Computer Science and Software Engineering,Vol.3,pages 107-114,April -2013.

[2] Er.Nikhil Chaturvedi and Er.Anand Rajavat An Improvement in K-means clustering Algorithm using Better Time and Accuracy International Journal of programming language and applications Vol.3,No.4,pages 13-19,October-2013.

[3] K.P.N.V.Satya sree and Dr.J V R Murthy ”Clustering based on cosine similarity measure” international journal of engineering science and advanced technology,Vol-2,508-512,Issue 3,April -2012.

[4] Gowtham S.,Bipul Syam Purkayastha An approach for document Preprocessing and K-Means Algorith Implementation 2014 4th International Conference on Advances in Computing and Communications ,pages 162-166,2014 IEEE.

[5] Suphakit Niwattanakul,Jatsada Singthongchai,Ekkachai Naenudorn and Supachanun Wanapu Using Of Jaccard Coefficient for Keywords Similarity Proceeding of the International Multiconference of Engineers and computer scientists2013 Vol-1,IMECS 2013 ,Hong Kong,March 13.

[6] S.C. Punitha, R. Jayasree andDr. M. Punithavalli, Partition Document Clustering using Ontology Approach, Multimedia and Expo, 2013 International Conference on Computer Communication and Informatics (ICCCI -2013), Jan.04 06,pages 1-5, 2013.

[7] Ranjana Agrawal ,Madhura Phatak, A Novel Algorithm for Automatic Document Clustering, 3rd IEEE International Advance Computing Conference (IACC) ,pages 877 - 882,IEEE, 2013. Achieve Trust in MANET, (IJANS) Vol. 2, No. 2, April 2012 DOI : 10.5121/ijans.2012.2206 53.

[8] Rui Xu, GDonald Wunsch, Survey of Clustering Algorithms, IEEE transactions on neural networks, VOL. 16, NO. 3,pages 645 - 678, MAY 2005.

[9] Shameem,efficient k-means algorithm integrated with jaccard distance measure for Document clustering.IEEE,2009.

[10] http://www.tfidf.com