An Investigation of Various Data Mining Based Clustering Techniques For Performing Clustering of Text Documents

Mr. Prince Agrawal

Rishiraj Institute of Technology(LNCT), RGPV University,

Sawer Road, Indore,India

Prof. Hemant Gupta

Rishiraj Institute of Technology(LNCT), RGPV University,

Sawer Road, Indore,India
prince.agrawal33@gmail.com
hemugupta3131@gmail.com

Abstract

Clustering means keeping similar objects together. Document clustering is an extension of clustering, which is related to keeping similar text documents together. Document clustering plays a vital role in development of search engines, where a group of document is required to listed as a result of query in minimum response time. This paper elaborates the concept of document cum text clustering. This paper presents a methodology for document clustering. This methodology is based on an efficient K means variant. This algorithm makes use of density based connected objects for selecting better clusters. It will result in overall improvement in clustering accuracy:

Keywords: *Document Clustering, Search Engine, Data Mining, Forecasting, Clustering Method.*

1. Introduction

The use of data mining [1,2] is placed in various decisions making task, using the analysis of the different properties and similarity in the different properties can help to make decisions for the different applications. Among them the prediction is one of the most essential applications of the data mining and machine learning. This work is dedicated to investigate about the decision making task using the data mining algorithms. Data mining is associated with extraction of non trivial data from a large and voluminous data set. Figure 1 shows the general working of data mining.

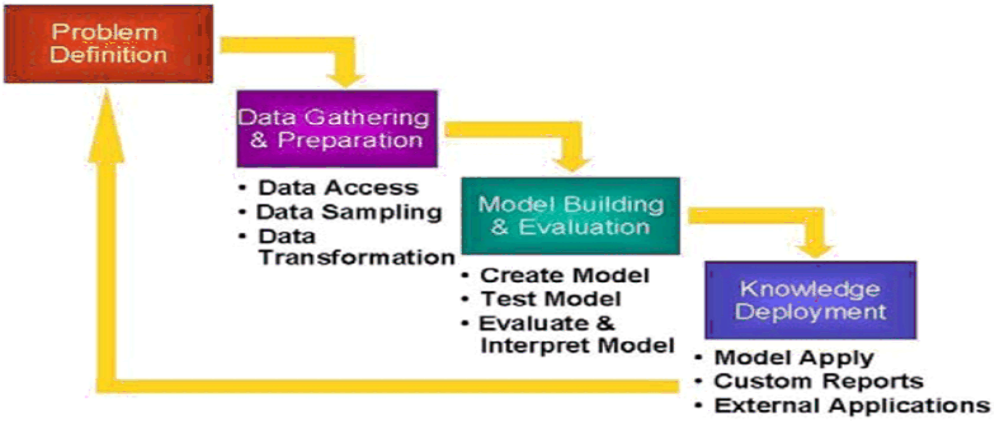

Figure 1: Data Mining

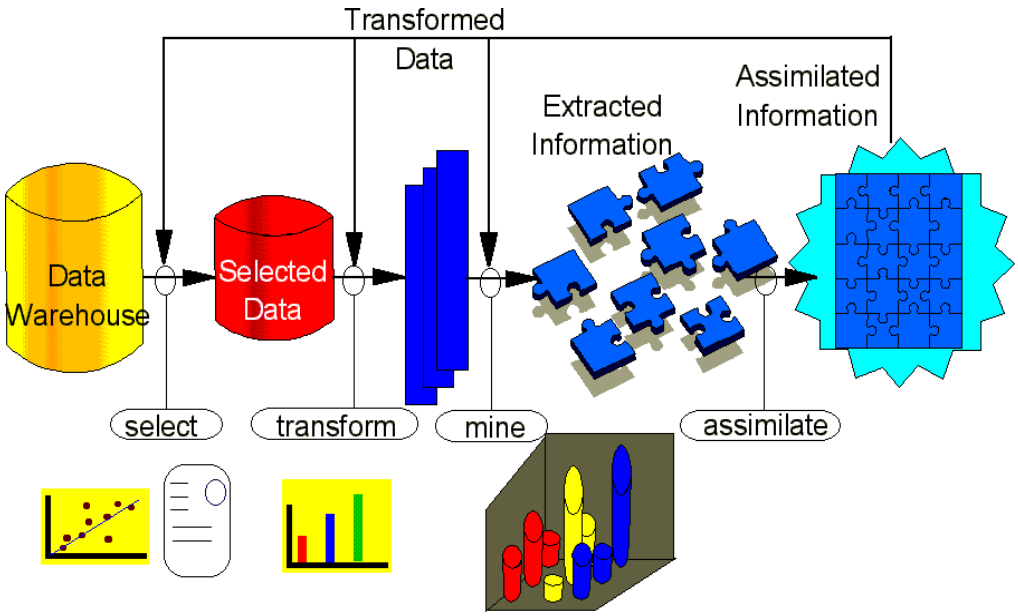

Figure 2: key steps in data mining

Figure 2 shows, key steps performed during the process of data mining. The data mining is a process of analysis of the data and extraction of the essential patterns from the data. These patterns are used with the different applications for making decision making and prediction related task. The decision making and prediction is performed on the basis of the learning of algorithms. The data mining algorithms supports both kinds of learning supervised and unsupervised. In unsupervised learning only the data is used for performing the learning and in supervised technique the data and the class labels both are required to perform the accurate training. In supervised learning the accuracy [3,4] is maintained by creating the feedbacks form the class labels and enhance the classification performance by reducing the error factors from the learning model.

Clustering is a partition of data into groups of related objects. Each set, called cluster, consists of objects which are similar to each other and dissimilar to the item of other groups. In other language, the principle of a high-quality document clustering approach is to decrease intra-cluster distances between documents. It is shown below in figure 2. In clustering is the allocation and the nature of information that will conclude cluster membership, in conflict to the classification where the classifier learn the association between objects and classes from a so set, i.e. a set of documents properly label by hand, and then replicates the learnt performance on unlabeled data

The document clustering framework is shown below in figure 3. Input are text documents. Then key words are identified in these documents. Then similarity is measured in these documents. Generally Euclidian distance is used as similarity measure. Then on basis of similarity documents are mapped in the correspondent clusters.

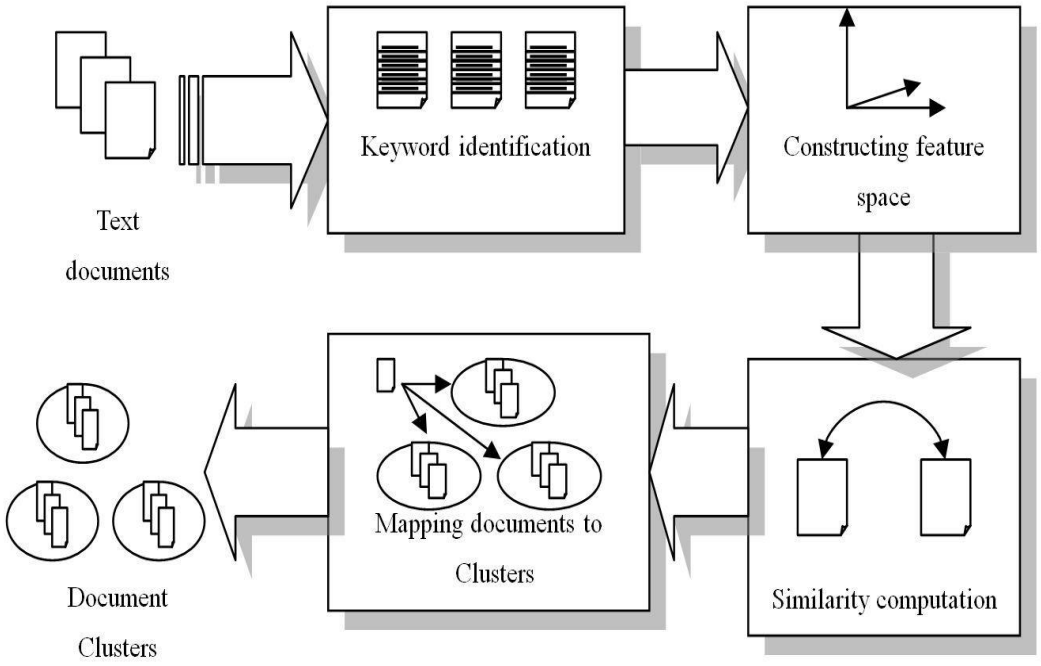

Figure 3: Document Clustering Framework[6]

2. Related Work:

[1] shows a correlated application domain of mining, e-mails are group by using structural, and domain-specific features. Three clustering methods (K-means, Bisecting K-means and EM) were used. [2] posited an way for clustering heterogeneous data streamswith uncertainty. [3]designed a new clustering approach by combination divisional and agglomerative clustering known as HPSO. It developed the cleverness of ants in a decentralized environment. This method proved to be very efficient as it performed clustering in a agglomerative manner . [4] define clustering-based scheme to recognize the fuzzy system. To start the mission, is tried to present a modular method, based on hybrid clustering method. [5] lent maintain to an incremental clustering for unqualified data using clustering collection. They initially compact unnecessary attributes if required, and then made use of accurate values of different attributes to form clustering memberships. [6] shows incorporated background for mining mails for forensic study, using classification and clustering method . [7] addressed the difficulty of clustering mails for forensic study where a Kernel-support variation of K-means was apply. The obtained outcome were examine personally, and the creator concluded that they are attractive and valuable from an analysis perspective. [8] The former, capable of maximize middling similarity within clusters and minimize the same among clusters, is a twosome similarity clustering. The latter attempt to generate approach from the manuscript, each technique representing one document set in particular. [9] mulled over a method about be short of software extracting method, which is a procedure of extracting information out of resource code. They offered a software extracting task with an integration of manuscript mining and link study technique.

[10] in organize to cluster the results from keyword searches. The underlying assumption is that the clustered results can increase the information retrieval efficiency, because it would not be required to review all the documents found by the client anymore. [11] shows (self-organize map) SOM-based algorithms used for clustering files with the aim of making the decision-making process achieved by the examiners more efficient. The files were clustered by taking into report their creation dates/times and their extensions. *[12]* a scribed data mining function and their various necessities on clustering procedure. The most important necessities considered are their potential to recognize clusters

implanted in subspaces. [13] predetermined iterative clustering method to evaluate preliminary cluster centers for K-means. This procedure is sufficient for clustering procedure for constant data.

3. Proposed Methodology

The outline of the proposed approach is as follows:

IMPLEMENTATION MODULES:

Ontology generation

- Pre-Processing element
- Count the amount of cluster

Ontology Clustering

- Clustering procedure

3.3.1 Preprocessing Module:

Stop-words- In preprocessing the first step is to remove inappropriate document metadata. A typical method to eradicate stop word is to judge against each term with a compilation of known stop words

Input:A document Data Base D and List of Stop words LD={d1, d2, d3… dk;

where1<=k<=i

Output: All legal branch manuscript term D **Algorithm**: For (every di in D) doFor(1 to j) do

Remove tij from di If (tij in list L) Remove tij from di

Word stemming- The development of suffix removal to general word stems. A stem is anatural group of words with similar meaning. The step of take away words to their original form, or stem. For illustration, the words “connected,” “connection”, “connections” are all reduced to the stem “connect.” Porter’s technique is the de facto standard stemming algorithm [3].

Algorithm:

Step 1: Gets rid of plurals and -ed or -ing suffixes

Step 2: Turns terminal y - i when there is a further vowel in the stem

Step 3: Maps double suffixes to single ones:-ization, -ational, etc.

Step 4: Deals with suffixes, -full, -ness etc.

Step 5: Takes off -ant, -ence, etc.

Step 6: Removes a final –e

Similarity Measurement- Several studies suggest that approx. 30% document in repository are similar, so checking the similarity document and removing it make our clustering technique to take less time. The method find simhash similarity take two influence doc[i],doc[j] as a parameter which are vector representation of document i and j and return the sim value i.e the similarity score which indicates the document are exactly similar or near similar

Similarity between two sets I & j is computed as follows:

Algorithm:

For i: =0 to N

For j: =0 to N

Sum worth :=((doc[i]*doc[j]))/Math.sqrt (doc[i]*doc[j])

Add _sim worth to the record Build_ matrix;

Next

Next

Where N is whole amount of data

Doc[i] for i=1, 2.....n are documents

3.3.2 Calculating the Clusters:

In this step, only the clusters are feed as input. It is stand for K.

3.3.3 Clustering Techniques:

For clustering the bunch of data taken from a newspaper articles we use K means & improved method so that comparison between this -

Steps K Means method:

Initialization- In the first tread data set, quantity of clusters and the center are defined for every cluster.

Classification- The similarity is calculated for all records point from the center and the data point having least similarity from the center of a set is assign to that picky cluster.

III. Center Recalculation- for the Clusters generated before, the center again evaluate means recalculation of the center

IV. Convergence Condition-

i. Stopping when getting a given or define number of steps.

Stopping when there no replace of data spot among data
iii. Stopping when a threshold rate is achieved.

If all of the above conditions are not satisfied, then apply step II and the whole process repeat again, until the given conditions are satisfied

But as described earlier there are problems and drawbacks in K-mean algorithm thus to overcome the problems a new method is being implemented on the data used the algorithm for the new improved technique is as follow

Steps of proposed and improved Clustering Technique

Output: D = {d1, d2, d3,..., di,..., dn }
di = { x1, x2, x3,..., xi,..., xm } k

Input: A value of k clusters.

Step 1. Select k=2 original cluster centers Ci randomly from data Xi .
Repeat following steps for every cluster centerStep 2. Find Euclidean similarity of each data objects Xi from cluster centers and allocate objects to cluster with least distance.

Step 3. Find min and max distance similarity beside with corresponding nearby aim and farthest object

Step 4. Evaluate two sets of items NPT and MPT enclose tightly joined objects to within distance: avg_dist= (Min_dist+Max_dist)/3

Step 5. Choose K i)NPTi ∩ MPTi=Φ ii) NPTi ∩NPTj=Φ and MPTi ∩MPTj=Φ If (i) legal then divide Ci and if both (i) and (ii) valid divide both center and assign new center as of corresponding cluster. If either condition is valid then goto step 2.

Step 6. Find mean for each cluster.

Step 7. If cluster value is same then exit.

The above Modified k-means algorithm has additional steps in traditional k-means for better cluster center selection. We use distance for transfer item to appropriate cluster by using these deliberate distance and we find near objects from c For selecting better cluster centers we use sets of densely connected objects. The NPT set enclose objects within avg_dist from min_obj and MPT set enclose objects within avg_dist distnace from max_obj

4. Result Analysis:

We implemented existing K-means algorithm and the proposed K means algorithm in Java. We have used news data set. The data set comprises We compared the performance of the k means & the proposed clustering technique. The sequence of execution is as follows:

- Removing Stop Words:
- Stemming
- Computing Term Frequency
- Distance Calculation
- Purity Checking

The results obtained are as follows:

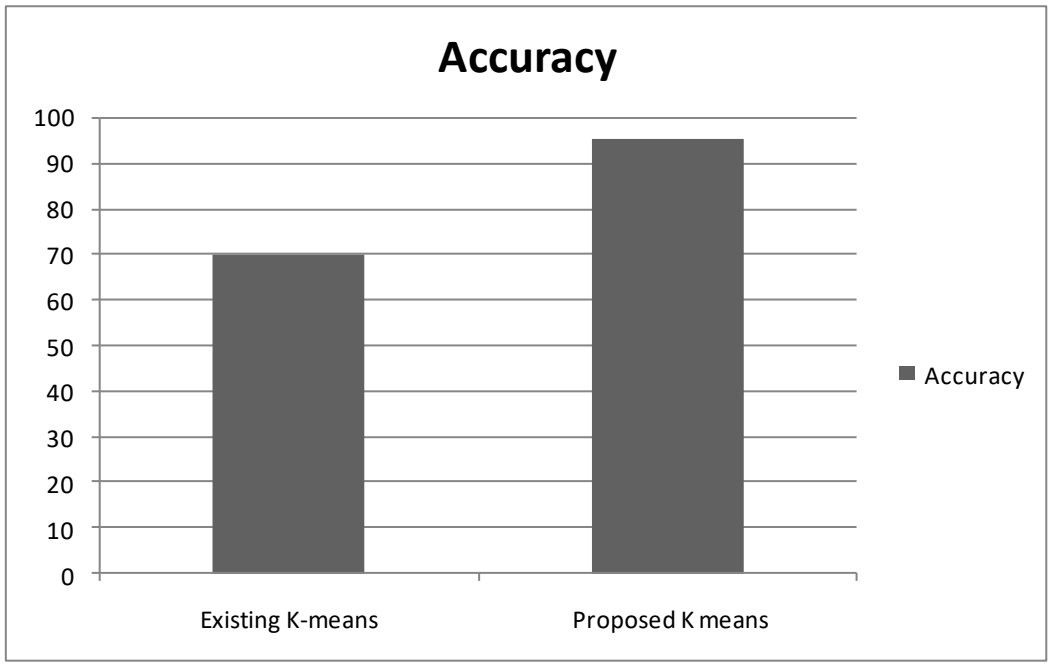

Figure : Accuracy Comparison

5. Conclusion:

In this paper, the focus is on Document Clustering which is very recent technology, we investigated many existing algorithms. As clustering plays a very vital role in various applications, many researches are still being done. The upcoming innovations are mainly due to the properties and the characteristics of existing methods. A methodology for document clustering has been proposed in this paper. This is based on the proposed K

means algorithm. The experimental results have shown that the accuracy of proposed method is better than the existing technique.

REFERENCES:

[1] F. Iqbal, H. Binsalleeh, B. C. M. Fung, and M. Debbabi, "Mining writeprints from anonymous e-mails for forensic investigation," *Digital Investigation, Elsevier*, vol. 7, no. 1–2, pp. 56–64, 2010.

[2] Guo-Yan Huang, Da-Peng Liang, Chang-Zhen Hu and Jia-Dong Ren, "An algorithm for clustering heterogeneous data streams with uncertainty", 2010 International Conference on Machine Learning and Cybernetics (ICMLC), Vol. 4, pp. 2059-2064, 2010.

[3] Alam, S., Dobbie, G., Riddle, P. and Naeem, M.A. "Particle Swarm Optimization Based Hierarchical Agglomerative Clustering", IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), Vol. 2, pp. 64-68, 2010.

[4] Shin-Jye Lee and Xiao-Jun Zeng, "A three-part input-output clustering-based approach to fuzzy system identification", 2010 10th International Conference on Intelligent Systems Design and Applications (ISDA), pp. 55-60, 2010.

[5] Li Taoying, Chne Yan, Qu Lili and Mu Xiangwei, "Incremental clustering for categorical data using clustering ensemble", 29th Chinese Control Conference (CCC), pp. 2519-2524, 2010.

[6] R. Hadjidj, M. Debbabi, H. Lounis, F. Iqbal, A. Szporer, and D. Benredjem, "Towards an integrated e-mail forensic analysis framework," *Digital Investigation, Elsevier*, vol. 5, no. 3–4, pp. 124–137, 2009.

[7] S. Decherchi, S. Tacconi, J. Redi, A. Leoncini, F. Sangiacomo, and R. Zunino, "Manuscript clustering for digital forensics analysis," *Computat. Intell. Security Inf. Syst.*, vol. 63, pp. 29–36, 2009

[8] Pallav Roxy and Durga Toshniwal, "Clustering Unstructured Manuscript Documents Using Fading Function", International Journal of Information and Mathematical Sciences, Vol. 5, No. 3, pp. 149-156, 2009

[9] Miha Grcar, Marko Grobelnik and Dunja Mladenic, "Using Manuscript Mining and Link Analysis for Software Mining", Lecture Notes in Computer Science, Vol. 4944, pp. 1-12, 2008.

[10] N. L. Beebe and J. G. Clark, "Digital forensic manuscript string searching: Improving information retrieval effectiveness by thematically clustering search results," *Digital Investigation,Elsevier*, vol. 4, no. 1, pp. 49–54, 2007.

[11] B.K.L.Fei, J.H.P.Eloff, H.S.Venter, and M.S.Oliver, "Exploring forensic data with self-organizing maps," in *Proc. IFIP Int. Conf. Digital Forensics*, 2005, pp. 113–123.

[12] . Aggarwal,C.C.Charu,andC.X.Zhai, Eds. "Chapter4: ASurvey of Manuscript Clustering Algorithms," in *Mining Manuscript Data*.NewYork: Springer, 2012.

[13] Shehroz S. Khan and Amir Ahmad, "Cluster Center Initialization Algorithm for K-means Clustering", Pattern Recognition Letters, Vol. 25, No. 11, pp. 1293-1302, 2004