A Comparative Study Of Document Clustering

Bandana Parida[1], Dr. Rakesh Chandra Balabantaray[2]

[1] [ME CSE (KE), PG Dept. of Computer Science and Application, Utkal University, Vani Vihar, Bhubaneswar,

[2] [Asst. Professor, Computer Science, IIIT, Bhubaneswar,]

Abstract: Data mining or knowledge discovery means extracting the knowledge or data from large amount of knowledge or data and summarising it into useful information. Data mining software has many tools for analysing data and summarising it. One of the tool is weka .It contains many machine learning algorithms. In this paper we are studying various clustering algorithms for the documents by using weka. Clustering means collecting a set of documents into group called clusters so that the documents in the same cluster are more similar than to other clusters.

Key-Words: Data mining algorithms, Weka tools, clustering algorithm .

1. Introduction:

Knowledge discovering process consists of different steps: Data Cleaning, Data Integration, Data Selection, Data Transformation, Data Mining, Pattern Evaluation and Knowledge Presentation. Data mining is a process in which many methods are applied to find out the data which are stored in database or data ware house. Data mining functionalities are: characterisation and discrimination, mining frequent pattern, association, correlation, classification and prediction, cluster analysis, outlier analysis and evolution analysis [1].
Three types of Data Mining techniques are Regression, Classification and Clustering. Clustering means, taking the similar documents into a cluster and other into another cluster. Clustering is an important technique for statistical data analysis including machine learning, pattern recognition, information retrial and bioinformatics. Here we are using weka data mining tool for clustering the documents. Then we are applying the stemming process to each clustering algorithm and finding out the difference between all algorithms that means how the documents are changing their cluster or group by applying the stemming algorithm.

2. Weka

Weka is one of the open source data mining software tool developed by University of Wail Kato in New Zealand that provides solution to many algorithms. Weka or Wooden (Gallirallus australis) is an endemic bird of New Zealand. It is a collection of machine learning algorithms for data mining tasks and only a tool kit such wide spread adaption and survive for an extended period of time [2]. WEKA is open source software issued under the GNU General Public License [3]. It is platform independent.

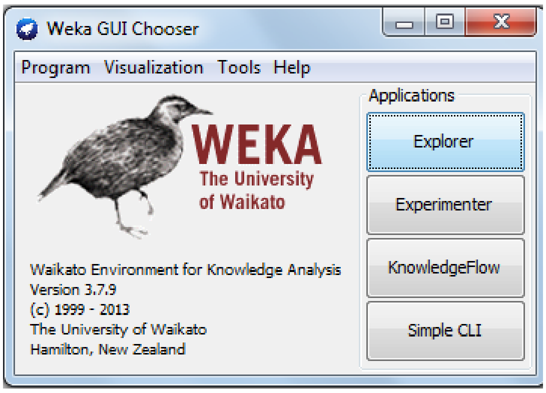

Figure1. View of weka tool

The GUI Chooser consists of four buttons:

- Explorer: It is an environment to explore the data with WEKA.
- Experimenter: It is an environment to perform experiments and conduct statistical tests between learning schemes.
- Knowledge Flow: The function of this environment is same as the Explorer but with a dragand-drop interface with an advantage of incremental learning support.
- Simple CLI: It provides a simple command-line interface which allows direct execution of WEKA commands for operating systems in which own command line interface is not provided.

When we click the “explorer button” we find Weka Explorer pre-processing, classification, clustering, association, attribute selection and visualisation tools. We have open the files which are must be in “.arff” format. Then we apply the clustering algorithms to all the documents.

3. Clustering Methods

a) Cobweb Clustering.
b) Expectation Maximization Clustering.
c) Farthest Fast Clustering.
d) Filtered Clustering.
e) Hierarchical Clustering.
f) Make Density Based Clustering.
g) Simple K-Means Clustering.

3.1 Cobweb Clustering

This algorithm is developed by machine learning researchers in 1980[4]. It provides cluster without any predefined number of clusters. Here each cluster is represented by probabilistically with a conditional probability. It uses an evaluation function called category utility to guide the construction of the tree.

a. This algorithm starts with an empty root node.
b. Instances are added one after another.
c. For each instance following options are taken.
    - The instance is classified into an existing class
    - A new class is created and the instance is placed into it
    - Two classes are combined into a single class (merging) and the new instance is placed in the resulting hierarchy;
    - A class is divided into two classes (splitting) and the new instance is placed in the resulting hierarchy.

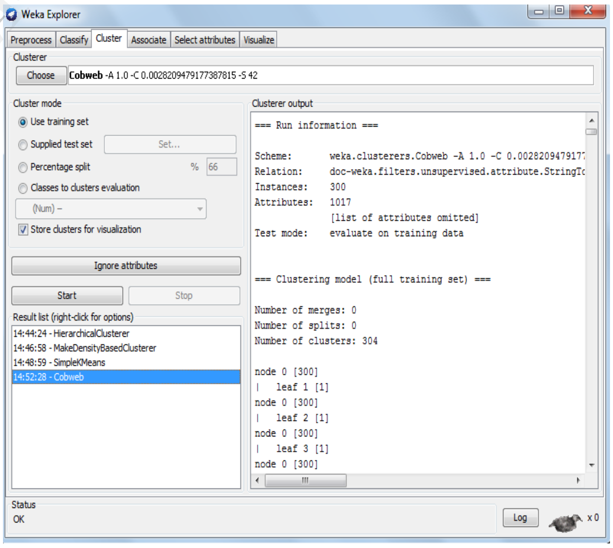

Figure 2: Cobweb clustering algorithm

3.2 Expectation Maximization Clustering

Expectation means computing the probability that each datum (attribute) is a member of each class (cluster), Maximisation means altering the parameters of each class (cluster) to maximise the probabilities [5]. It is convergence but not necessarily correct.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 6, June - 2013

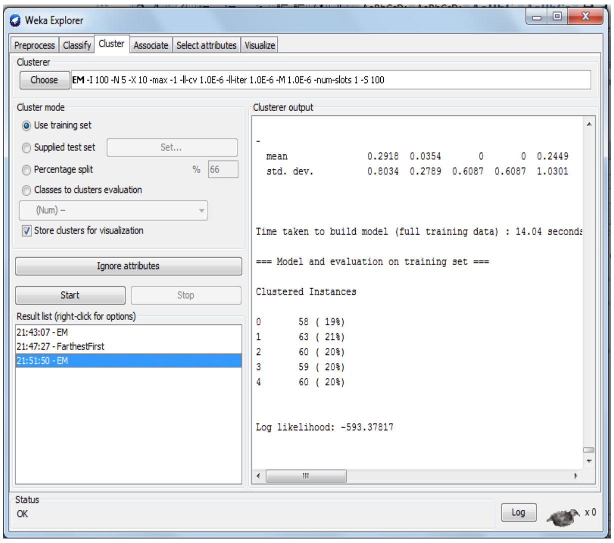

Figure 3: EM clustering algorithm

3.3 farthest Fast Clustering

This algorithm is developed by Hochbaum and Shomoy in 1985: A best possible heuristic for K-centre problem [6]. It is a variant of K means that places each cluster centre in turn at the point farthest from existing cluster centre.

By taking the TF and IDF the following analysis of the documents are shown below:-

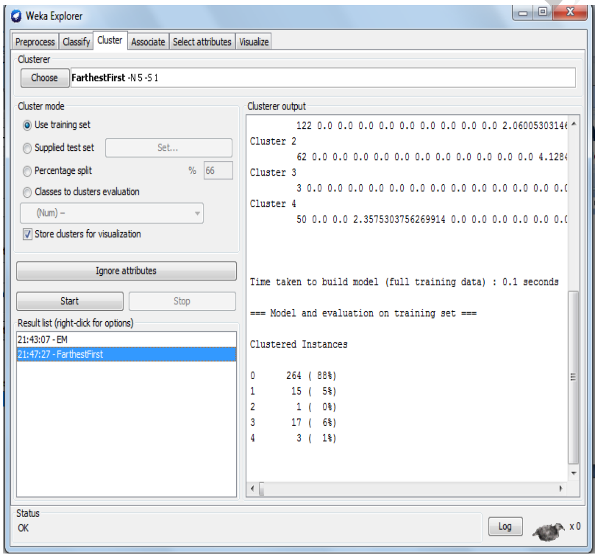

Figure 4: Farthest fast clustering algorithm

To find out the result of the algorithm we right click on the visualise cluster assignment, a new window is opened and show the result in the form of a graph. By clicking the "save button" we can save the result in the form of "arff." Format.

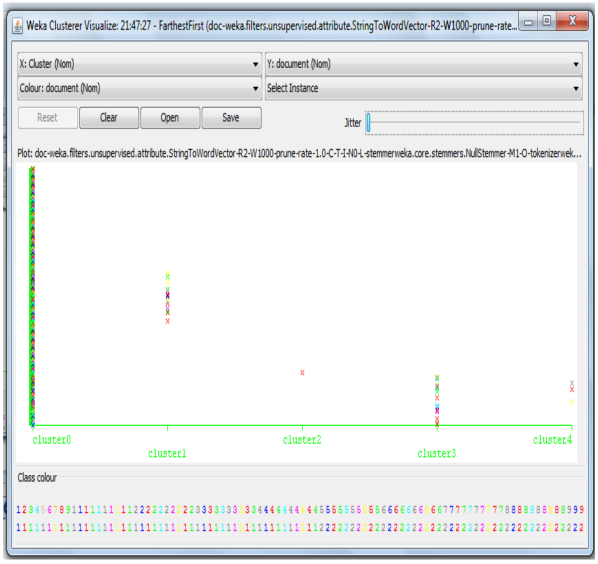

Figure 5: Result of Farthest fast in form of graph

3.4 Hierarchical Clustering

Here the cluster is generated hierarchically that means a tree of clusters called as dendrograms[7]. It is of two types.

a) Agglomerative (bottom up)

- Start with 1 point (singleton).
- Recursively add two or more appropriate clusters.
- Stop the process when k number of clusters is achieved.

b) Divisive (top down)

- Start with a big cluster.
- Recursively divided into smaller clusters.
- Stop the process when k number of clusters is achieved.

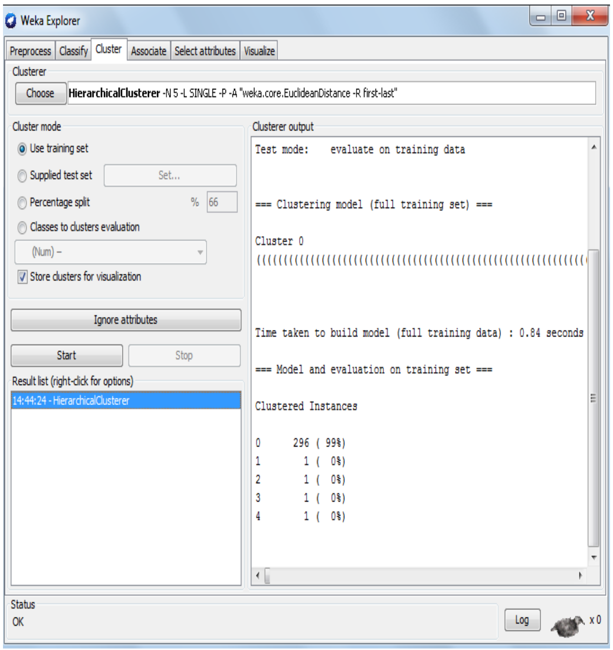

Figure 6: Hierarchical clustering algorithm

3.5 Make Density Based Clustering

This algorithm is proposed by Martin Ester, Hans-Peter Kriegel, Jorge Sander and Xiaowei Xu in 1996.In this algorithm we try to find the cluster according to the density of data point in a region. The main idea of this clustering is for each of cluster the neighbourhood of given radius has contain at least minimum number of instances. DBSCAN [8] is the most common clustering algorithm and also most cited scientific literature.

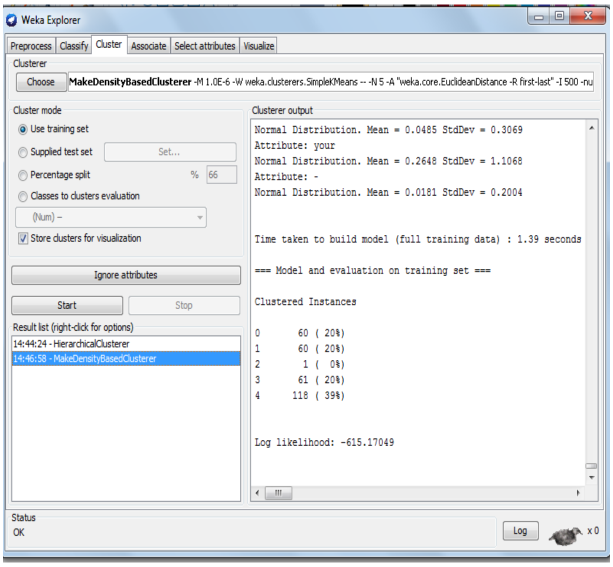

Figure 7: Make density based clustering algorithm

3.6 Filtered Clusterer

It is a class for running an arbitrary cluster on data that has been passed through an arbitrary filter. Filtering is the process of removing special characters and punctuation that are not required for providing the result.

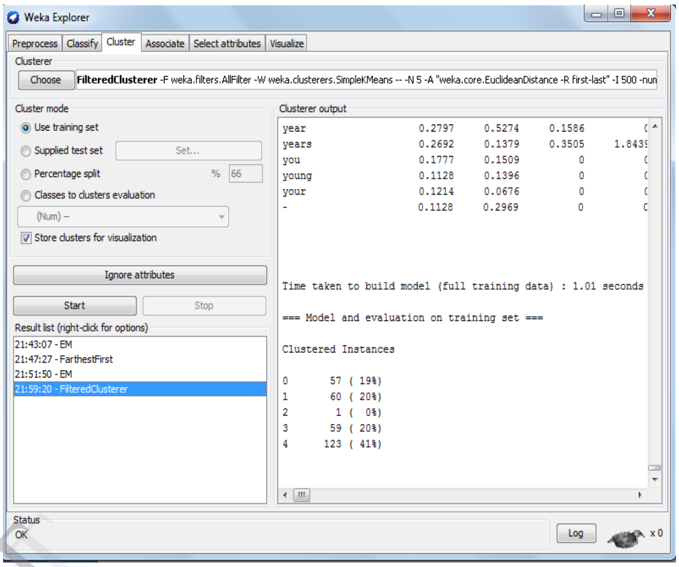

Figure 8: Filtered clusterer algorithm

3.7 Simple K-Means Clustering

The term "k means" was first used by James Macqueen in 1967 [9] is the one of the unsupervised learning algorithm and it was developed by Stuart Lloyd in 1957 based on the technique of pulse-code modulation. The aim of the algorithm is partitioning n documents [10] into k clusters in which each document belongs to the cluster with the nearest means. It provides an output which is most efficient in terms of execution time.

The algorithm is worked in the following steps [11]:

1. Arbitrarily choose k points from data set D as initial cluster. These points represent the initial group of centroids.
2. Assign the object to the cluster or group which has closest centroid.
3. Recalculate the position of the k-centroids.
4. Repeat step 2 and 3 until the centroids are no longer change.

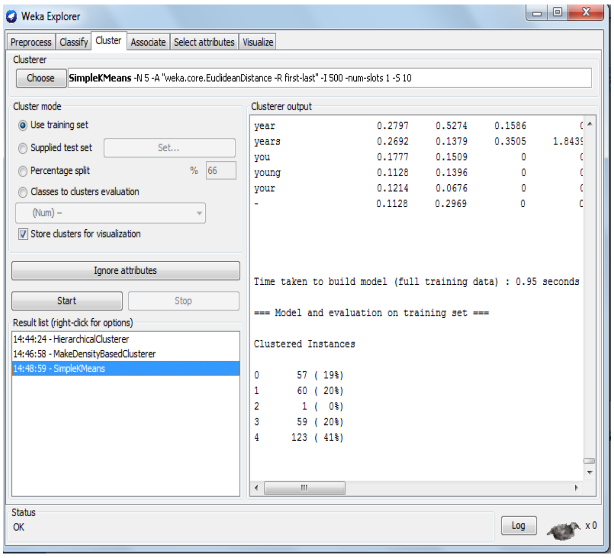

Figure 9: Simple k-means clustering algorithm

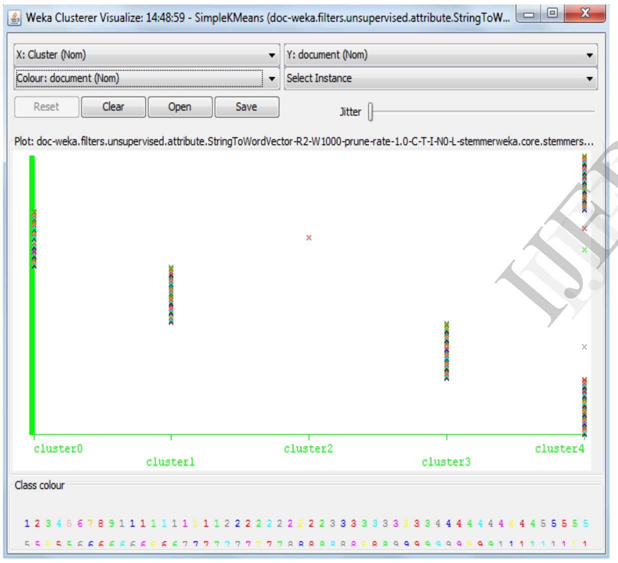

Figure10: Result of Simple k-means in form of graph

4. Comparison

The above section involves all the clustering algorithms using weka tool. We are taking 300 numbers of documents which are from five domains like Pollution, Entertainment, Constitution of India, Festivals of India and Indian History. Then we make a comparative study of the documents by taking all clustering algorithms using weka tool.

This comparative study involves three cases:

1. By taking both term frequency transform (TF) and inverse document frequency transform (IDF)
2. By taking only term frequency transform (TF)
3. By taking Stemmer(Lovins Stemmer & Snowball Stemmer) with term frequency transform (TF) and inverse document frequency transform (IDF)

The results of the clustering algorithms are shown in four different tables:

Table -1(Comparison result by taking both TF and IDF):

Name	No. of clusters	Cluster instances	Time taken to build model	Un-clustered Instances
EM	5	0:58(19%) 1:63(21%) 2:60(20%) 3:59(20%) 4:60(20%)	13.64 seconds	0
Farthest Fast	5	0:264(88%) 1:15(5%) 2:1(0%) 3:17(6%) 4:3(1%)	0.05 seconds	0
Filtered Cluster	5	0:57(19%) 1:60(20%) 2:1(0%) 3:59(20%) 4:123(41%)	1 seconds	0
Hierarchical Clustering	5	0:296(99%) 1:1(0%) 2: 1(0%) 3: 1(0%) 4: 1(0%)	0.91 seconds	0
Density based Clustering	5	0:60(20%) 1: 60(20%) 2:1(0%) 3:61(20%) 4:118(39%)	1.2 seconds	0
K-Means	5	0:57(19%) 1:60(20%) 2:1(0%) 3:59(20%) 4:123(41%)	0.95 seconds	0

Table 2 (Comparison result by taking only TF):

Name	No. of clusters	Cluster instances	Time taken to build model	Un-clustered Instances
EM	5	0:58(19%) 1:60(20%) 2:63(21%) 3:60(20%) 4:59(20%)	15.18 seconds	0
Farthest Fast	5	0:263(88%) 1:16(5%) 2:1(0%) 3:17(6%) 4:3(1%)	0.06 seconds	0
Filtered Cluster	5	0:57(19%) 1:60(20%) 2:1(0%) 3:59(20%) 4:123(41%)	1.12 seconds	0
Hierarchical Clustering	5	0:296(99%) 1:1(0%) 2: 1(0%) 3: 1(0%) 4: 1(0%)	0.73 seconds	0
Density based Clustering	5	0:59(20%) 1:60(20%) 2:1 (0%) 3:61(20%) 4:119(40%)	1.36 seconds	0
K-Means	5	0:58(19%) 1:63(21%) 2:60(20%) 3:59(20%) 4:60(20%)	1.3 seconds	0

Table 3 Comparison result by taking snowball stemmer with TF and IDF:

Name	No. of clusters	Cluster instances	Time taken to build model	Un-clustered Instances
EM	5	0:58(19%) 1:63(21%) 2:60(20%) 3:59(20%) 4:60(20%)	13.95 seconds	0
Farthest Fast	5	0:264(88%) 1:15(5%) 2:1(0%) 3:17(6%) 4:3(1%)	0.03 seconds	0
Filtered Cluster	5	0:57(19%) 1:60(20%) 2:1(0%) 3:59(20%) 4:123(41%)	0.97seconds	0
Hierarchical clustering	5	0:296(99%) 1:1(0%) 2: 1(0%) 3: 1(0%) 4: 1(0%)	0.73seconds	0
Density based clustering	5	0:60(20%) 1:60(20%) 2:1(0%) 3:61(20%) 4:118(39%)	1.2 seconds	0
K-Means	5	0:57(19%) 1:60(20%) 2:1(0%) 3:59(20%) 4:123(41%)	0.98 seconds	0

Table 4 comparison result by taking lovins stemmer with TF and IDF:

Name	No. of clusters	Cluster instances	Time taken to build model	Un-clustered Instances
EM	5	0:58(19%) 1:63(21%) 2: 60(20%) 3: 59(20%) 4: 60(20%)	13.96 seconds	0
Farthest Fast	5	0:239(80%) 1:2(1%) 2: 21(7%) 3: 32(11%) 4: 6(2%)	0.09 seconds	0
Filtered Cluster	5	0:114(38%) 1:60(20%) 2: 3(1%) 3: 59(20%) 4: 4(2%)	0.97 seconds	0
Hierarchical Clustering	5	0:296(99%) 1:1(0%) 2: 1(0%) 3: 1(0%) 4: 1(0%)	0.48 seconds	0
Density based Clustering	5	0:111(37%) 1:60(20%) 2: 4(1%) 3: 60(20%) 4: 65(22%)	1.12 seconds	0
K-Means	5	0:114(38%) 1:60(20%) 2: 3(1%) 3: 59(20%) 4: 64(21%)	0.89 seconds	0

4. Conclusion

In this paper we have projected various clustering algorithms in document clustering using weka. We do not require deep knowledge about algorithms when working with weka. So weka is more suitable data mining tool. We found that the k-means clustering algorithm is simplest and provide better performance as compared to other algorithms while taking the above three cases. But when the time factor is concerned, the farthest fast clustering algorithm executes faster & EM clustering algorithm takes more time than all other algorithms. Hierarchical clustering algorithm is more sensitive for noisy data than other algorithms. We also found that density based clustering algorithm is not suitable for data with high variance in density.

References:

[1]Han J. and Kamber M.: “Data Mining: Concepts and Techniques,” Morgan Kaufmann Publishers, San Francisco, 2000.

[2]Sapna Jain, M Afshar Aalam and M N Doja, “K-means clustering using weka interface”, Proceedings of the 4th National Conference; INDIACom-2010.

[3] Weka Manchine Learning Project, http://www.cs.waikato.ac.nz/~ml/index.html.

[4] Narendra Sharma , Aman Bajpai , Mr. Ratnesh Litoriya, “Comparison the various clustering algorithms of weka tools”, International Journal of Emerging Technology and Advanced Engineering ,(ISSN 2250-2459, Volume 2, Issue 5, May 2012).

[5]Jurmug and net\project\misc\em

[6] E.B Fawlkes and C.L. Mallows. A method for comparing two hierarchical clustering. Journal of the American Statistical Association, 78:553–584, 1983.

[7] Manish Verma, Mauly Srivastava, Neha Chack, Atul Kumar Diswar and Nidhi Gupta, “A Comparative Study of Various Clustering Algorithms in Data Mining”, International Journal of Engineering Research and Applications (IJERA) Vol. 2, Issue 3, May-Jun 2012, pp.1379-1384.

[8] Slava Kisilevich, Florian Mansmann, Daniel Keim ―P-DBSCAN: A density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos, University of Konstanz

[9] MacQueen J. B., "Some Methods for classification and Analysis of Multivariate Observations", Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability.University of California Press. 1967, pp. 281–297.

[10] Jinxin Gao, David B. Hitchcock ―James-Stein Shrinkage to Improve K-meansCluster Analysis‖ University of South Carolina,Department of Statistics November 30, 2009

[11] Jiawei Han and Micheline Kamber, “Data Mining: Concepts and Techniques”, Morgan Kaufmann Publishers,second Edition, (2006).