Clustering Technique on Search Engine Dataset using Data Mining Tool

MD. Ezaz Ahmed
Department of Computer Science and Engineering
itm University
Gurgaon, Haryana, India-122017
ezazahmed@itmindia.edu

Preeti Bansal
Department of Computer Science and Engineering
itm University
Gurgaon, Haryana, India-122017
preeti.bansal7@gmail.com

***Abstract*- Unlabeled document collections are becoming increasingly common and mining such databases becomes a major challenge. It is a major issue to retrieve good websites from the larger collections of websites. As the number of available Web pages grows, it is become more difficult for users finding documents relevant to their interests. Clustering is the classification of a data set into subsets (clusters), so that the data in each subset share some common trait - often proximity according to some defined distance measure. By clustering we improve the quality of websites by grouping similar websites in groups. This paper addresses the applications of data mining tool Weka by applying k means clustering to find clusters from huge data sets and find the attributes that govern optimization of search engines.**

***Keywords*—Dataset; Websites; Data mining; Weka; k-means**

I. INTRODUCTION

The Web has experienced continuous growth since its creation. As of March 2002, the largest search engine contained approximately 968 million indexed pages in its database. Finding the right information from such a large collection is extremely difficult [3].Information extraction plays a vital role in today's life. How efficiently and effectively the relevant documents are extracted from World Wide Web is a challenging issue. As today's search engine does just string matching, documents retrieved may not be so relevant according to user's query. By clustering the websites, the websites having values of attributes are in particular range are grouped together [6]. Data is collected from various websites source code like their title length, number of keywords in title, url length, number of backlinks etc and based on this we derive the conclusion. A popular technique for clustering is based on K-means such that the data is partitioned into K clusters. In this method, the groups are identified by a set of points that are called the cluster centers. This paper is organized as follows:

Section 1 is introduction and describes the problem and tool used. Section 2 discuss the algorithm used and describes the data mining technique adopted Section 3 describes the tool and dataset and concludes the paper.

II. DOCUMENT CLUSTERING

Document clustering analysis plays an important role in document mining research. A widely adopted definition of optimal clustering is a partitioning that minimizes distances within a cluster and maximizes distances between clusters.

In this approach the clusters and, to a limited degree, relationships between clusters are derived automatically from the documents to be clustered, and the documents are subsequently assigned to those clusters [1].

Users are known to have difficulties in dealing with information retrieval search outputs especially if the outputs are above a certain size. Clustering can enable them to find the relevant documents more easily and also help them to form an understanding of the different facets of the query that have been provided for their inspection. This project aimed to investigate the websites that are in top 5 in one cluster and other sites in second cluster. Clustering based on k-means is closely related to a number of other clustering and location problems. These include the Euclidean k-medians , in which the objective is to minimize the sum of distances to the nearest center, and the geometric k-center problem in which the objective is to minimize the maximum distance from every point to its closest center.

A. K means Algorithm

K-Means clustering is a very popular algorithm to find the clustering in dataset by iterative computations. It has the advantages of simple implementing and finding at least local optimal clustering. K-Means algorithm is employed to find the clustering in dataset. The algorithm [2],[9] is composed of the following steps:

- Initialize k cluster centers to be seed points. (These centers can be randomly produced or use other ways to generate).
- For each sample, find the nearest cluster center, put the sample in this cluster and recompute centers of the altered cluster (Repeat n times).
- Exam all samples again and put each one in the cluster identified with the nearest center (don't recompute any cluster centers). If members of each cluster haven't been changed, stop. If changed, go to step 2.

III. ORGANIZATION OF DATA

It's important for search engine to maintain a high quality websites. This will improve the optimization. We made a database in which following attributes we take length of title, keywords in title, Domain length, and number of backlinks and Top rank website .

A. Working with Weka on Dataset

Open Weka, and then click on right side option explorer then Open data file under preprocess option which is in csv or arff format [4],[5]. As we choose the explorer option it will appear as given below, the screen shot in fig 1. Clearly indicate the open file option. Now we click on view open file and choose the data set. Weka provides filters to accomplish all of these preprocessing tasks, they are not necessary for clustering in Weka .This is because Weka

978-0-7695-4941-5/13 $26.00 © 2013 IEEE
DOI 10.1109/ACCT.2013.15

Simple K Means algorithm automatically handles a mixture of categorical and numerical attributes. This algorithm automatically normalizes numerical attributes when doing distance computations [7].

This gives all attributes that are present in dataset. We can select any one which we want to include or select all.

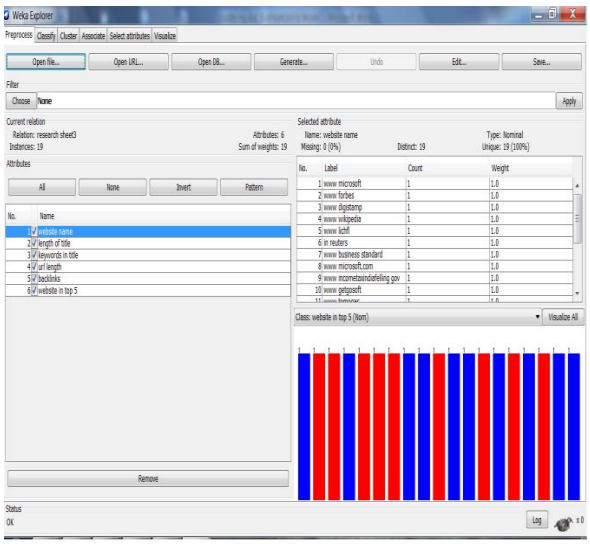

Fig.1: Opening page

After this just click on cluster tab and click on choose button on left side and select clustering algorithm which we want to apply, we select simple k means the screen appears below in fig 2. [8]

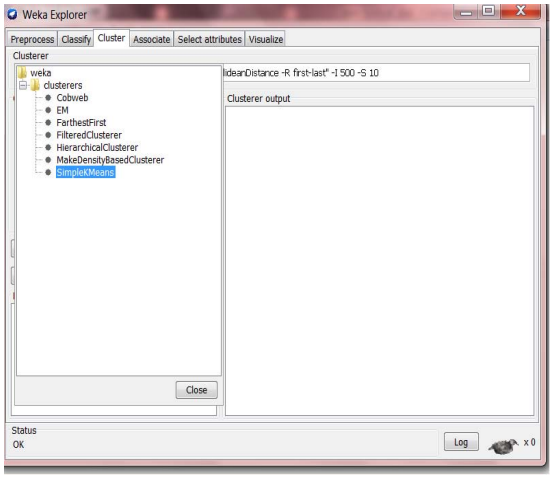

Fig.2: Select algorithm

Next, click on the text box to the right of the "Choose" button to get the pop-up window shown in Fig 3, for editing the clustering parameter. In the pop-up window we enter **2** as the number of clusters and we leave the value of "seed" as is. The seed value is used in generating a random number which is, in turn, used for making the initial assignment of instances to clusters.

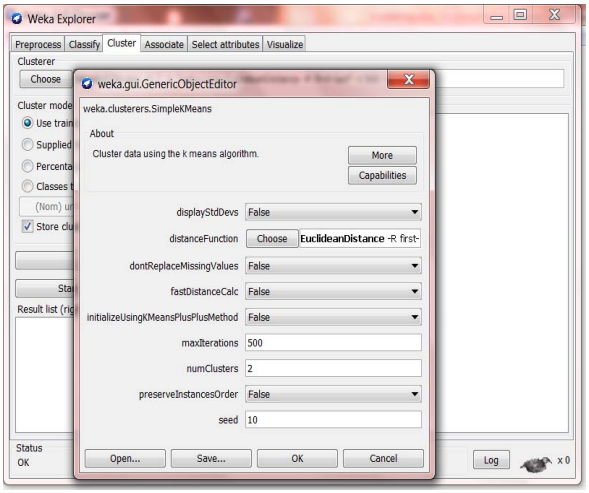

Fig 3: Choose parameters

Note that, in general, K-means is quite sensitive to how clusters are initially assigned. Thus, it is often necessary to try different values and evaluate the results. [10].

Once the options have been specified, we can run the clustering algorithm. Here we make sure that in the "Cluster Mode" panel, the "Use training set" option is selected, and we click "Start". We can right click the result set in the "Result list" panel and view the results of clustering in a separate window**.**

The result window shows the centroid of each cluster as well as statistics on the number and percentage of instances assigned to different clusters. Cluster centroids are the mean vectors for each cluster (so, each dimension value in the centroid represents the mean value for that dimension in the cluster). Thus, centroids can be used to characterize the clusters.

The result shows that in cluster 0 there are 13 websites that have length of title 59 characters long, keywords in title are 5, url length 22 characters long and number of backlinks are 6638 and in cluster 1 there are 16 websites that have length of title 36 characters long, keywords in title are 3, url length 29 characters long and number of backlinks are 19163 as shown in fig 4.

Another way of understanding the characteristics of each cluster is through visualization. We can do this by right-clicking the result set on the left "Result list" panel and selecting "Visualize cluster assignments". This pops up the visualization window as shown in Fig 5.

In this, we choose the cluster number and any of the other attributes for each of the three different dimensions available (x-axis, y-axis, and color). Different combinations of choices will result in a visual rendering of different relationships within each cluster.

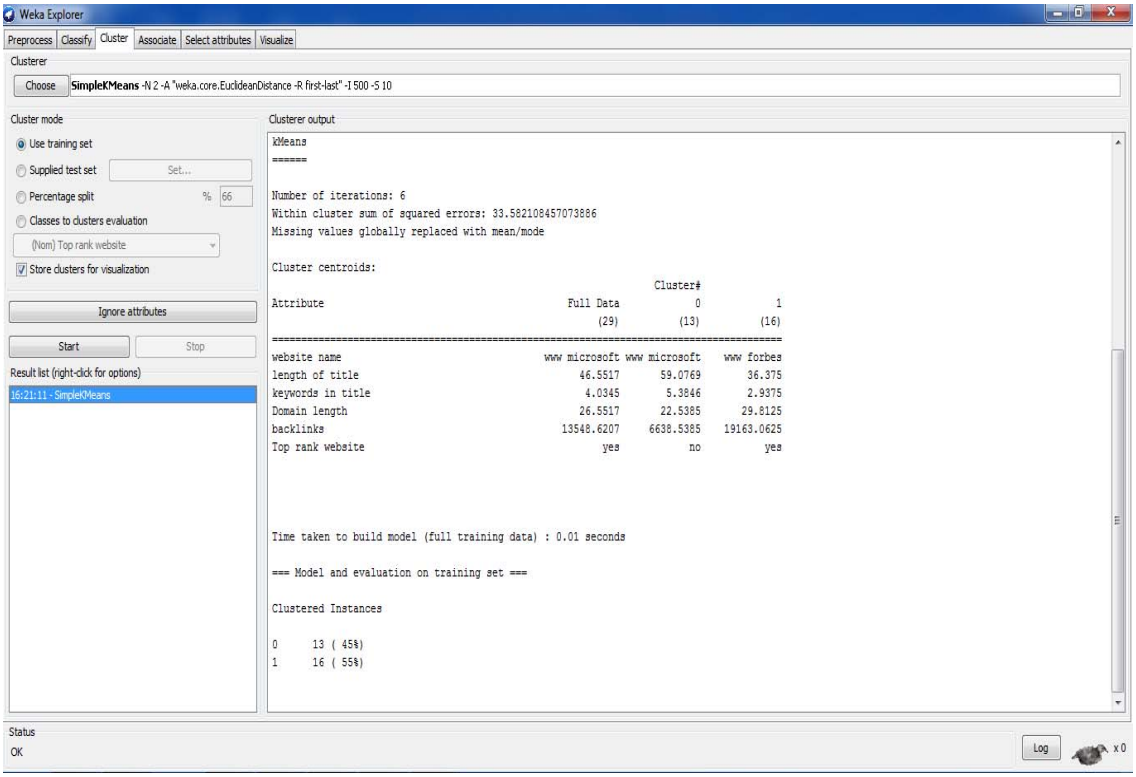

Fig. 4: Result of clustering

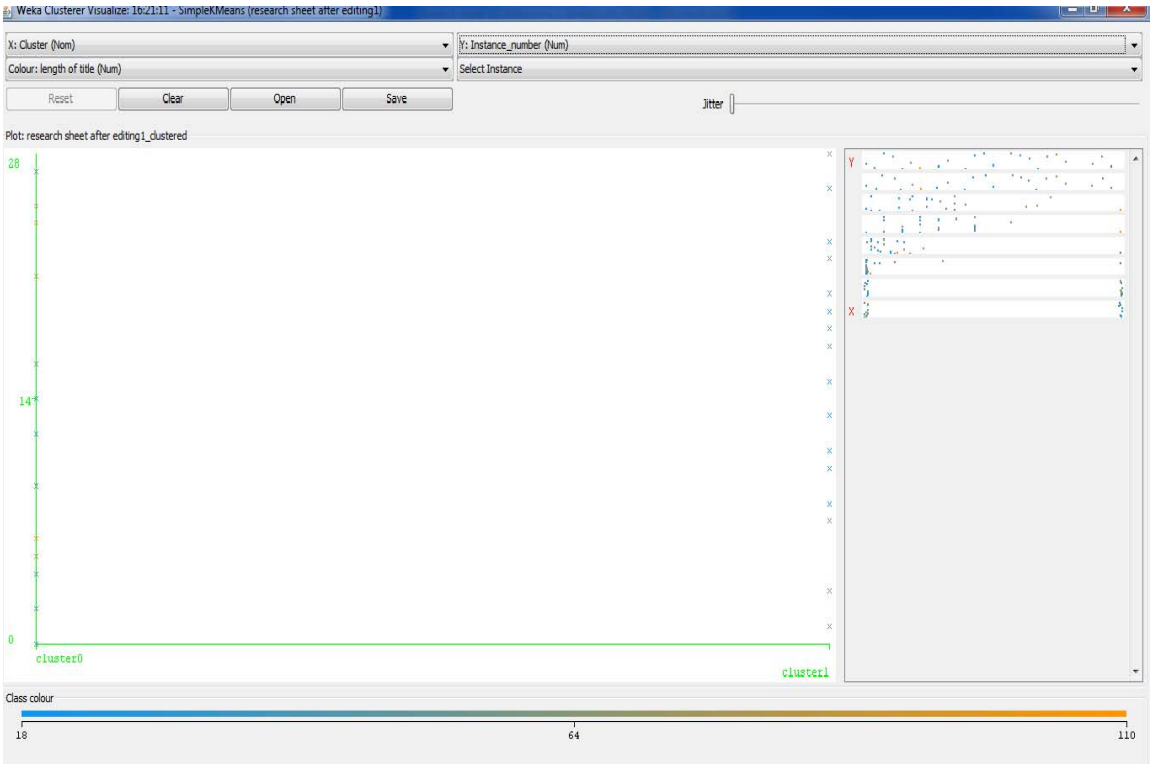

Fig.5: Visual result

In the above example, we have chosen the cluster number as the x-axis, the instance number (assigned by Weka) as the y-axis, and the "length of title" attribute as the color dimension. This will result in a visualization of the distribution of length of title in two clusters.

As more and more data is collected from websites we can get more detail and can find attributes as by this method we find backlinks > 19000 , length of title < 40 , keywords in title > 3 and Domain length < 30 is good for search engine optimization.

REFERENCES

[1] A Document Clustering Algorithm for Web Search Engine Retrieval System, Hongwei Yang School of Software Yunnan University, Kunming 650021, China;

[2] S. Kantabutra, Efficient Representation of Cluster Structure in Large Data Sets, Ph.D. Thesis, Tufts University, Medford MA, September 2001.

[3] Wang Jun, OuYang Zheng-Zheng "The Research of K-Means Clustering Algorithm Based on Association Rules "

[4]. http://maya.cs.depaul.edu/classes/ect584/weka/k-means.html

[5]. http://www.cs.ccsu.edu/~markov/weka-tutorial.pdf.

[6] http://thesai.org/Downloads/Volume3No4 /Paper_20 Knowledge_Discovery_in_Health_Care_Datasets Using_Data_Mining_Tools.pdf

[7] www.gtbit.org/downloads/dwdmsem6/dwdmsem6lman.pdf

[8] http://www.iasri.res.in/ebook/win_school_aa/notes/WEKA.pdf

[9] R. Kannan, S. Vempala, and Adrian Vetta, "On Clusterings Good, Bad, and Spectral", Proc. of the 41st Foundations of Computer Science, Redondo Beach, 2000.5.

[10] http://www.bvicam.ac.in/news/INDIACom%202010%20 Proceedings / papers/Group3/INDIACom10_388_Paper.pdf.