International Journal of Information and Computation Technology.
ISSN 0974-2239 Volume 3, Number 6 (2013), pp. 505-510

http://www. irphouse.com /ijict.htm

K-means Clustering Technique on Search Engine Dataset using Data Mining Tool

Minky Jindal[1] and Nisha Kharb[2]

[1,2]CSE/IT Department, ITM University, Sector-23A, Gurgaon, INDIA.

Abstract

Today World Wide Web is growing wide as its name. It seems a bit difficult to manage all the information with search engine, as new modules have to be added into it. For that purpose, a technique name clustering comes in to everyone's mind. Clustering provided an organized way to manage our search engine. It is no less than a coal which runs the engine. This paper addresses the applications of data mining tool Weka by applying k means clustering to find the clusters from huge data sets and clustering that provide a building hand in the optimization of search engine.

Keywords: Component; Dataset, Data mining, k-means, Weka

1. Introduction

With the rapid development of Internet, society has entered the Internet age. According to incomplete statistics, the number of Internet users in China in 2008 reached 200 million people, 89% of those Internet users mainly use the Internet to obtain information, of which 88.8% will make use of Internet search engines search information, second only to send and receive E-Mail. Web search engines have become an important part [3].

With the rapid growth of web pages, it is very tough for users to find the relevant document of their interests. By applying clustering, data is collected from various websites source code like their title length, number of keywords, URL length, number of back links, in links .Based on these parameters clusters are made to derive the conclusion. A well-known technique for clustering is based on K-means, in which the data is partitioned into K clusters known as cluster centers after which it can enable

users to find the information to the point providing user interaction with the search output.

2. Original K-means Algorithm

From a practical point of view, clustering analysis is one of the main tasks of data mining. It is now used in many areas like knowledge discovery, pattern recognition and so on. Many clustering analysis algorithm are available of which the most well-known is the K-means algorithm which is based on division. Clustering can enable users to find the relevant documents more easily. This paper aimed to investigate the websites that are in top in one cluster and other sites in second cluster and for top ranking we need URL, back-links, in-links, length of title are required. "Clustering based on k-means" that it is closely related to a number of other clustering and location problems which include the Euclidean k-medians which minimize the sum of distances to the nearest center, and the geometric k-center problem, which aimed to minimize the maximum distance from every point to its closest center [1].

A. K means Algorithm

K-Means clustering is a very popular algorithm to find the clusters in a dataset by iterative computations. It has the advantage of simple implementation 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:

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

3. Organization of Data

It's important for search engine to maintain a high quality websites. A database is to be made in which following attributes should be there like take length of title, keywords in title, number of back-links, in-links, URL length, ranking WEKA and Tanagra a data mining tool can be used to extract meaningful knowledge from large set of data. We derive the conclusion by taking data collected from various websites source code like their title length, number of keywords in title, URL length, number of back links, in-links etc.

3.1 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 arff or csv format [4], [5]. As we choose the explorer option it will appear as shown in the screen shot in fig 1. Open the file option tab Now 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 'Simple K Means' algorithm automatically handles a mixture of categorical and numerical attributes. This algorithm automatically normalizes numerical attributes when doing distance computations is there [8].This gives all the attributes that are present in the dataset. We can select any one which we want to include or select all.

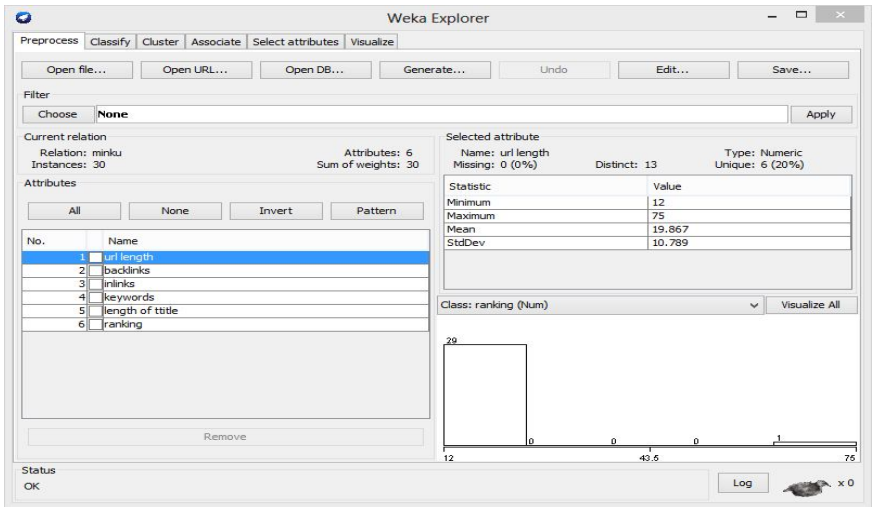

Fig. 1: Open a CSV file.

After this, 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 which is shown in fig 2. [7]

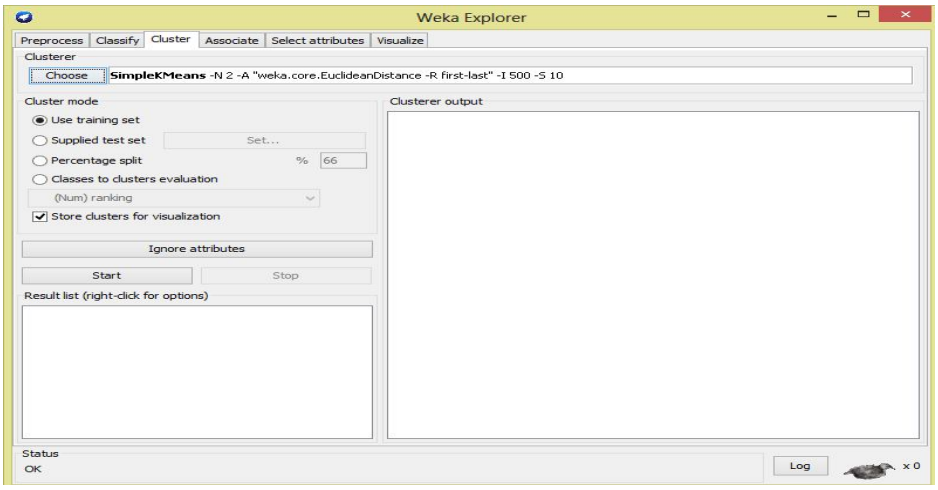

Fig. 2: Select Algorithm.

Next, click on the text box to the right of the "Choose" button to get the pop-up window as shown in Fig 3, for editing the clustering parameter. In the pop-up window, enter 4 as the number of clusters.

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 evaluation of result should be done [10].

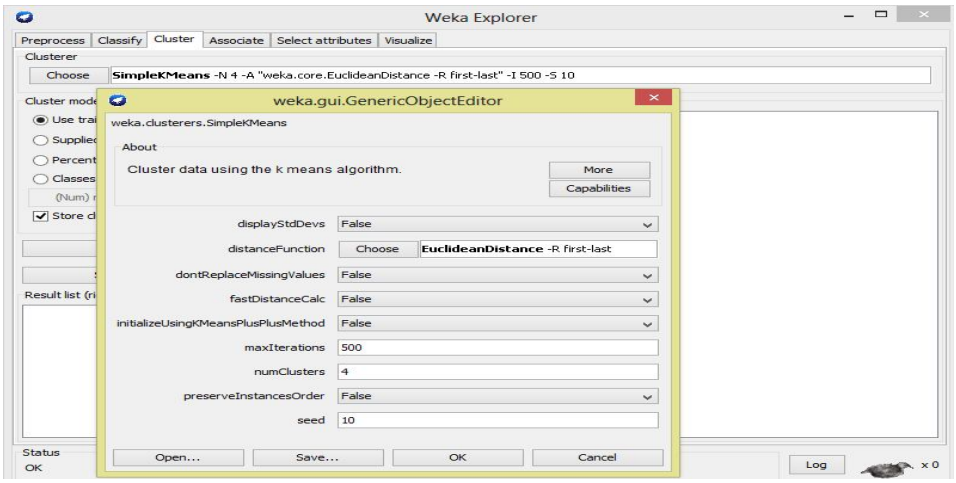

Figure 3: Choose Attributes.

Once the options have been specified, run the clustering algorithm. Make sure that in the "Cluster Mode" panel, the "Use training set" option is selected, and then click "Start". If user wants to view the results of clustering in a separate window then right click the result set in "Result List" panel. This result window will show 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 cluster

The result shows that in cluster 0 there are 7 websites that have URL length 18 characters long, back-links 1965857, in-links 682, keywords 15 characters long and length of title 108 and in cluster 1 there are 1 websites that have URL length 19characters long, back-links 69700, in-links 733, keywords 12 characters long and length of title 19 and in cluster 2 there are 1 websites that have URL length 15 characters long, back-links 605000, in-links 1041, keywords 7 characters long and length of title 15 and in cluster 3 there are 21 websites that have URL length 20 characters long, back-links 233239, in-links 272, keywords 16 characters long and length of title 106 as shown in fig 4.

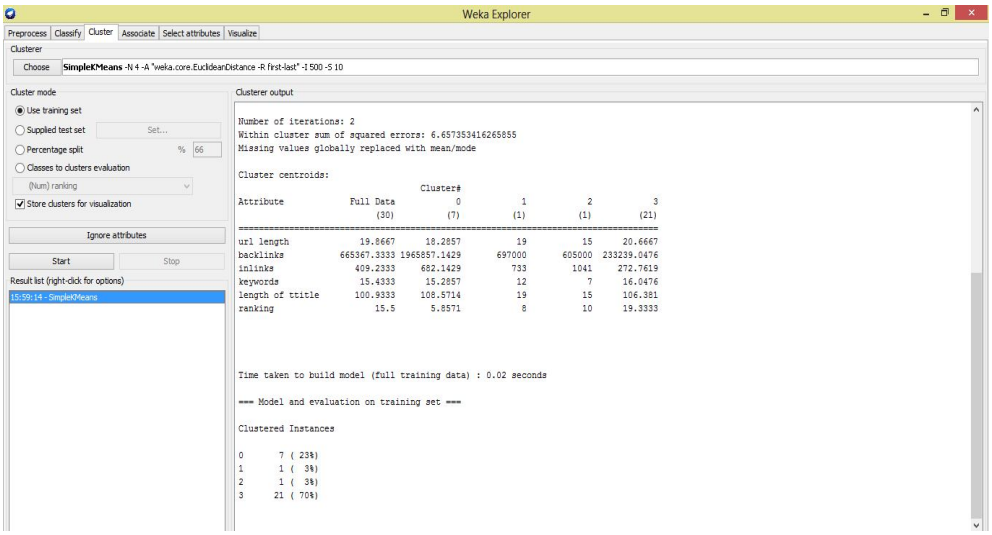

Figure 4: Result of K-MEANS Clustering.

Another way of understanding the characteristics of each cluster is through visualization. This can be done 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, 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.

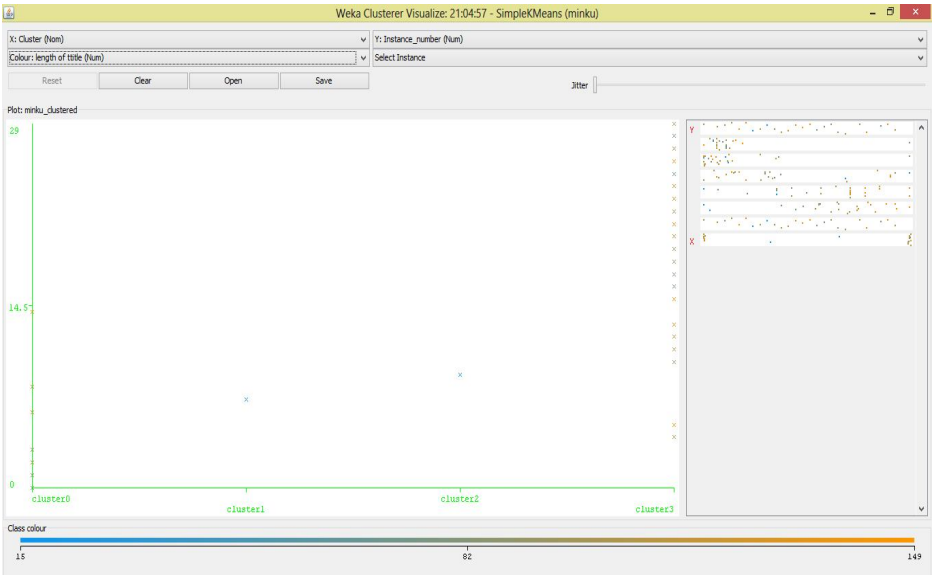

Figure 5: Visual result.

In the above example, choose 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.

4. Conclusion

As more and more data is collected from websites we can get more detail and can find attributes as by this method we find back-links > 60000 , length of title < 50 , keywords in title > 3 and URL length < 25 and in-links >200 is good for search engine optimization.

References

[1] Hongwei," A Document Clustering Algorithm for Web Search Engine Retrieval System", In proceeding of International Conference on e-Education, e-Business, e-Management and e-Learning,2010,PP 383-386, DOI 10.1109/IC4E.2010.

[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 ",In proceeding International Conference on Challenges in Environmental Science and Computer Engineering,2010,PP 285-286, DOI 10.1109/CESCE.2010.26.

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

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

[9] R. Kannan, S. Vempala, and Adrian Vetta, "On ClusteringsGood, Bad, and Spectral",In Proc. of the 41st Foundations of Computer Science, Redondo Beach, 2005.

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