Analysis of Social Networking Sites Using K- Mean Clustering Algorithm

[1]D. S. RAJPUT, [2]R. S. THAKUR, [3]G. S. THAKUR & [4]NEERAJ SAHU \

[1,2&3]Department of Computer Applications, MANIT, Bhopal (MP), India
[4]Singhania University Rajasthan
E-Mail : Dharm_raj85@yahoo.co.in[1], Ramthakur2000@yahoo.com[2],
ghanshyamthakur@gmail.com[3], neerajsahu79@gmail.com[4]

Abstract -Clustering is one of the very important technique used for classification of large dataset and widely applied to many applications including analysis of social networking sites, aircraft accidental, company performance etc. In recent days, Communication, advertising through social networking sites are most popular and interactive strategy among the users. This research attempts to find the large scale measurement study and analysis, effectiveness of communication strategy, analyzing the information about the usage, people's interest in social network sites in promoting and advertising their brand in social networking sites. The significance of the proposed work is determined with the help of various surveys, and from people who use these sites. Further a more specific pre-processing method is applied to clean data and perform the clustering method to generate patterns that will be work as heuristics for designing more effective social networking sites.

Keywords *- Knowledge discovery, K-mean clustering, pre-processing, nearest neighbour searching, social networking sites.*

I. INTRODUCTION

We begin with a brief overview of social networks sites. Internet is primarily a source of communication, information and entertainment. In recent years; the use of social networking sites has been increasing. The use of these sites, interaction between the people is becoming easy. It is used by school colleges and IT professionals etc. It is important to understand why people use these websites; some people use them for business purposes, find new deals, legal and criminal investigations etc. Few social networking sites such as Facebook(2004), Twitter (2006), Myspace (2003), Orkut (2004), Friendster (2002), hi5 (2003), Google+(2011) etc, where people are connected with others directly.

Cluster analysis is a popular statistical tool for finding groups of respondents, objects, or cases that are similar to one another but different from those in other groups [1,2,11]. Analysis of social networking sites is closely dependent on clustering algorithms. There are many existing clustering algorithms such as K-Means, Fuzzy C-Means (FCM), CLERA, PAM, CLERANS etc [1,2,3,8,10,15] have their own pros and cons. K-Means is very fast but its center value is dependent on the initial assumptions K-means clustering (k-means), simply speaking, is an algorithm to classify or to group objects based on attributes or features into k number of group [2,5,13,14,16]. The grouping is done by minimizing the distances between data and the corresponding cluster centroid. In the application of means, we need to decide the value of k before starting the program, it should be noticed that different value of k will cause different levels of accuracy of the grouping.

II. LITERATURE REVIEW

Tapas Kanungo at el. in 2002 proposed An Efficient k-Means

Clustering Algorithm: Analysis and Implementation [5]. This paper presents a simple and efficient implementation of Lloyd's k-means clustering algorithm, which is called filtering algorithm. This algorithm is easy to implement, requiring a kd-tree as the only major data structure.

Gengxin Chen at el. in 2003 proposed Evaluation and Comparison of Clustering Algorithms in Analyzing ES Cell Gene Expression Data [6]. This paper given embryonic stems cell gene expression data, and applied several indices to evaluate the performance of clustering algorithms. This study may provide a guideline on how to select suitable clustering algorithms and it may help raise relevant issues in the extraction of meaningful biological information from microarray expression data.

Xu Yang at el. in 2010 proposed K-Means Based Clustering on Mobile Usage for Social Network Analysis Purpose [4]. This work studied data mining for social network analysis purpose, which aims at finding people's social network patterns by analyzing the information about their mobile phone usage.

Xiaoyan Li at el. in 2011 proposed Hybrid Retention Strategy Formulation in Telecom Based on k-means Clustering Analysis [7]. This paper proposed an idea of formulating hybrid strategy to retain valuable customers using clustering technology.

Objectives of the study:-

I. To analyze easily the various conditions responsible for the various social networking site used by the youth.

II. To analyze the great help at which networking site is mostly used.

III. To analyze the effective communication strategy through social networking sites.

IV. To study the effectiveness of brand communication through social networking sites from its users and communicators.

V. To find the impact of interaction through these communication among Indian.

III. PROPOSED METHODOLOGY

In this proposed work Analysis of social networking sites is totally dependent on clustering algorithms. The existing clustering algorithm K-mean is very fast algorithm. This algorithm is used to classify features into k number of group or to group objects based on attributes. The grouping is done by minimizing the similarity between data and the corresponding cluster centroid.

This section presents data collection, data preprocessing and clustering methodology to generate clusters. There are many friends and relatives who are on facebook, Skype, Google+, Orkut etc. First open the friend's page and save the HTML document. This procedure repeat for user's profile. Finally we have documents available for classification in unstructured format.

The fig.1 shows proposed framework which consists three modules.

a) Text pre-processing module

b) Clustering algorithm module.

c) Cluster extraction and Specific result module. This framework will receive input from unstructured HTML data. The first module will perform pre-processing and extract document set D, and second module will perform K- mean clustering technique to generate clusters. Finally the last module provides results analysis.

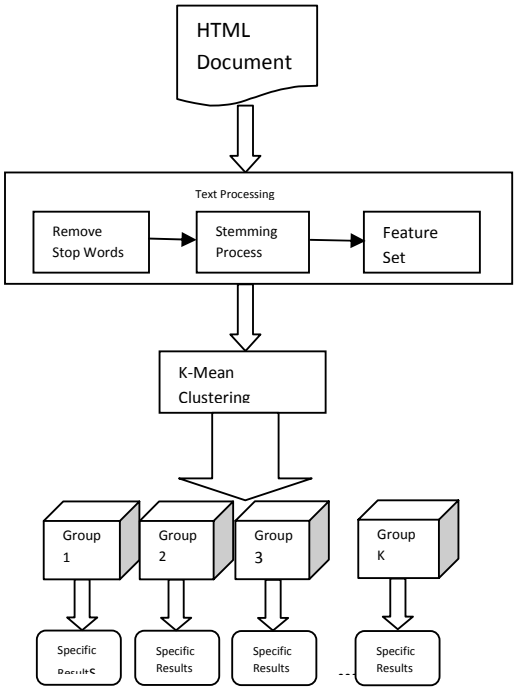

Figure 1: Framework for HTML Document Clustering

For the K-mean algorithm we have to decide value of K when beginning algorithm starts, it is noticed that different value of k will cause different levels of accuracy of the grouping[2,3,5,13,17]. The basic steps of K-mean clustering algorithm are:

Input: Number of cluster K.

Preprocessed Dataset.

1. Start
2. {
3. Take k samples from total number of N randomly as the centroid of each cluster.
4. Now Calculate the D of the remaining N-k sample to each centroid, and assign them to the cluster with the nearest centroid.
5. }
6. After each assignment, again calculate the centroid of the attainment cluster.
7. Now go to step 2 until find no new assignment.
8. Stop

IV. EXPERIMENTAL RESULTS

The proposed approach algorithm is applied in a social networking site dataset to generate clusters.

To explore the behavior of database we have choose the fields (Age Group, Time-spend, sex, occupation, Type of social site etc.). Aiming at the social site uses habits of Indian people; we have collected the data from the survey (like Family, Friends circle, college students etc.). All the experiments for finding the clustering results are performed on Pentium 2.6 GHz Processor, 2 GB RAM, Microsoft Windows 7.

The input means plots found in the Cluster node Results display the input means for the variables that were used in the clustering analysis over all of the clusters. The input means are normalized using a scale transformation function:

$$Y=(X-\min(X))/ (\max(x))-(\min(x))$$

For example, assume 5 input variables $Y_i= Y_1,...,Y_5$ and 3 clusters C_1, C_2, and C_3. Let the input mean for variable Y_i in cluster C_j be represented by M_{ij}. Then the normalized mean, or input mean, SMij becomes

$$SM_{ij}=(M_{ij}-\min(M_{i1},M_{i2},M_{i3})/((\max(M_{i1},M_{i2},M_{i3}))-(\min(M_{i1},M_{i2},M_{i3})))$$

The initial seeds must be complete cases, that is, training cases that have no missing values, and are required to be separated by a Euclidean distance that is of at least the value specified for the Minimum distance between cluster seeds (radius). By default, the initial seeds are chosen to be as far apart as possible; that is, seed replacement is set to full.

Using the data of social networking, the reproducibilities were as follows, for Initial seeds:

Maxclusters=40 Maxiter=1 and it is shown by Table.1

Cluster	1	2	3	4	5	6	7	8
1	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000
4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000
5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000
6	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000
7	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
8	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 1: Initial seeds

After first iteration we find the cluster Summary that has 8 clusters and find their frequency and distance between cluster centroids, which is shown by Table.2.

The Statistics tab of the Clustering Results Browser displays a table of clustering statistics Table.3 produced by the node's underlying vector quantization procedure. The entire training data set is used to calculate the following statistics for each cluster.

The root-mean-square standard deviation (RMS STD) measures the homogeneity of the cluster formed at any given step. It essentially measures the compactness or homogeneity of a cluster. Clusters in which consumers are very close to the centroid are compact clusters. The smaller the RMS STD, the more homogeneous or compact is the cluster formed at a given step. A large value of RMS STD suggests that the cluster obtained at a given step is not homogeneous, and is probably formed by merging of two very heterogeneous clusters. the first cluster would have a low RMSSTD whereas the second cluster would have a high RMSSTD. Notice that the cluster with a low RMSTD is relatively more homogeneous than the cluster with a high RMSTD. In general a cluster solution with a low RMSTD is preferred as it implies that the resulting clusters are homogeneous.

Cluster	Frequency	Root Mean Square Std Deviation	Maximum Distance from seed to Observation	Radius Exceeded	Nearest cluster	Distance between cluster centroids (mean)
1	33	0	0	-	8	1.4142
2	11	0	0	-	8	1.4142
3	15	0	0	-	8	1.4142
4	9	0	0	-	8	1.4142
5	10	0	0	-	8	1.4142
6	8	0	0	-	8	1.4142
7	6	0	0	-	8	1.4142
8	7	0	0	-	7	1.4142

Table 2: Cluster Summary after First Iteration

Variable	Total Standard Deviation	Within STD	R-Square	RSQ/(1-RSQ)
1	0.47380	0	1.000000	.
2	0.25764	0	1.000000	.
3	0.31587	0	1.000000	.
4	0.27393	0	1.000000	.
5	0.23982	0	1.000000	.
6	0.36037	0	1.000000	.
7	0.28894	0	1.000000	.
8	0.30288	0	1.000000	.
OVER ALL	0.32177	0	1.000000	.

Table 3: Statistics for Variables

*R-Square for predicting the variable from the cluster

*RSQ/(1-RSQ), which is the ratio of between-cluster variance to within cluster variance

Approximate Expected Over-All R-Squared = 0.52901

WARNING: The two values above are invalid for correlated variables.

In these results we find the cluster Mean. After final seed iteration Table.4 we find the final cluster Summary that has 6 clusters and find their frequency and distance between cluster centroids.

Cluster	Frequency	RMS Std Deviation	Maximum Distance from seed to Observation	Radius Exceeded	Nearest cluster	Distance between cluster centroids
1	46	0.2385	1.1375	-	2	1.2469
2	11	0	0	-	1	1.2469
3	15	0	0	-	1	1.2469
4	9	0	0	-	1	1.2469
5	10	0	0	-	1	1.2469
6	8	0	0	-	1	1.2469

Table 4: Cluster Summary after final Seed

A. Distance between clusters:-

It is shown in fig.2. The graph axes are determined from multidimensional scaling analysis, using a matrix of distances between cluster means as input. Therefore, it may appear that clusters overlap, but in fact, each case is assigned to only one cluster. The distance among the clusters is based on the criteria that are specified to construct the clusters. For illustrating clean the distance between clusters also shown by Table 5.

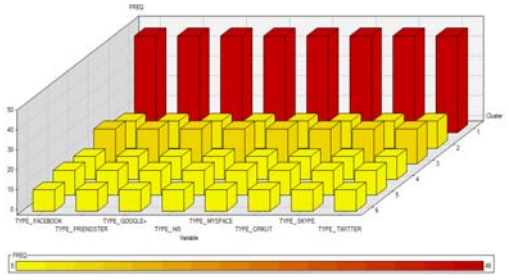

Figure 2: Distance between Clusters

Cluster	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6
1	0	1.2469	1.2469	1.2469	1.2469	1.2469
2	1.2469	0	1.2469	1.4142	1.4142	1.4142
3	1.2469	1.4142	0	1.4142	1.4142	1.4142
4	1.2469	1.4142	1.4142	0	1.4142	1.4142
5	1.2469	1.4142	1.4142	1.4142	0	1.4142
6	1.2469	1.4142	1.4142	1.4142	1.4142	0

Table5: Distance between Clusters

B. Cluster Tree:-

Cluster tree in fig.3 displays a decision tree (path) for selected cluster. The decision tree is based on the sample of the training data set that was configured in the Clustering node configuration interface, specifically in the Preliminary Training and Profiles subtab of the Data tab. The cluster variable is used as the target variable, and the tree enables us to identify influential inputs. The fig.3 shows the all node portion of the tree for all clusters.

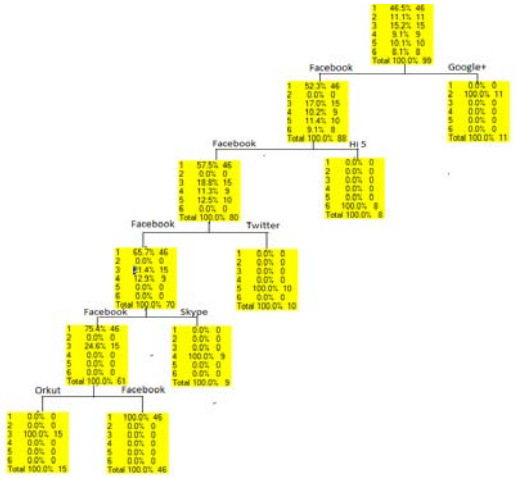

Figure 3: Clusters Tree

V. CONCLUSIONS

This paper analyzed social networking sites data using K-mean classification algorithm. The experimental results of text document classification on social networking sites dataset shows that 46% user preferred Facebook, 15% user preferred Orkut, 11% user preferred Google+, 10% user preferred Twitter, 9% user preferred Skype, 8% user preferred Hi5. This analysis concludes that the most common used website is the facebook.

1. By this analysis we can easily understand the various conditions responsible for the various social networking sites used by the youth.
2. The analysis is a great help at which networking site is mostly used.
3. This analysis also shows that this method works efficiently, for large text data.

ACKNOWLEDGMENT

This work is supported by research grant from MPCST, Bhopal M.P., India, Endt.No. 2427/ CST/R&D/2011 dated 22/09/2011.

REFERENCES

[1] Bryan Orme and Rich Johnson, Sawtooth Software, "Improving K-Means Cluster Analysis: Ensemble Analysis Instead of Highest Reproducibility Replicates" in 2008.

[2] Han I and Kamber M, "Data Mining concepts and Techniques,"Morgar Kaufmann Publishers,2000.

[3] D. Boley. Principal direction divisive partitioning. Data Mining and Knowledge Discovery, 2(4):325–344, 1998.

[4] Xu Yang, Yapeng Wang, Dan Wu, Athen Ma, "K-Means Based Clustering on Mobile Usage for Social Network Analysis Purpose" IEEE in 2010, pp 223-228.

[5] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, Angela Y. Wu, "An Efficient k-Means Clustering Algorithm: Analysis and Implementation" in IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 24, NO. 7, JULY 2002 ,pp 881-892.

[6] Gengxin Chen, Saied A. Jaradat, Nila Banerjee, Tetsuya S. Tanaka, "Evaluation and Comparison of Clustering Algorithms in Analyzing ES Cell Gene Expression Data" in 2003 pp 1-33.

[7] Xiaoyan Li, Yang Huang, Shujuan Li and Yishi Zhang, "Hybrid Retention Strategy Formulation in Telecom Based on k-means Clustering Analysis" IEEE in 2011, pp 978-981.

[8] Soon, M. C. , John, D. H., and Yanjun, L., "Text document clustering based on frequent word meaning sequences," Data & Knowledge Engineering, vol. 64, pp. 381-404, 2008.

[9] Chun-Ling Chen , Frank S.C. Tseng , Tyne Liang "An integration of WordNet and fuzzy association rule mining for multi-label document clustering" Data & Knowledge Engineering 69 (2010) pp 1208–1226 .

[10] Julie Beth Lovins, "Development of a Stemming Algorithm" Mechanical Translation and Computational Linguistics, vol.11, nos.1 and 2, March and June 1968.pp. 22-31.

[11] Larose, D. T., "Discovering Knowledge in Data: An Introduction to Data Mining", ISBN 0-471-66657-2, John Wiley & Sons, Inc, 2005. International Journal of Distributed and Parallel systems (IJDPS) Vol.1, No.1, September 2010.

[12] D.R. Recupero, "A new unsupervised method for document clustering by using WordNet lexical and conceptual relations, Information Retrieval" 10 (6) (2007) pp 563–579.

[13] Bryan Orme & Rich Johnson, "Improving K-Means Cluster Analysis: Ensemble Analysis Instead of Highest Reproducibility Replicates"in 2008.

[14] Andreas Hotho, Andreas N¨urnberger, Gerhard Paa, "A Brief Survey of Text Mining", in 2005.

[15] R.Vishnu Priya, A.Vadivel, R.S.Thakur, "Frequent Pattern Mining using Modified CP-Tree for Knowledge Discovery", Advanced Data Mining and Applications, LNCS-2010 volume 6440, pp. 254–261, © Springer-Verlag Berlin Heidelberg 2010

[16] R.S. Thakur, R.C. Jain, K.R. Pardasani, "Method of Conjugate Gradient: A Numerical Method for Mining Knowledge from Technical Data", in Research Hunt An International Multi Disciplinary Journal, Bhopal Vol. 1(1):182-189, 2006

[17] D.S. Rajput, R.S. Thakur, G.S. Thakur "Rule Generation from Textual Data by using Graph Based Approach" International journal of computer application, (IJCA) 0975 – 8887, New york USA, ISBN: 978-93-80865-11-8, Volume 31– No.9, October 2011.