A Study on Customer Segmentation Using K-Means Clustering for Online Shoppers

***Dr. R. Mary Metilda, **Mr. Vishnu Durai. R. S, ***Agarshana. P**

* Head of the Department, Sri Ramakrishna Engineering College, Coimbatore.
** Assistant Professor (Senior Grade), Sri Ramakrishna Engineering College, Coimbatore.
*** Student, Sri Ramakrishna Engineering College, Coimbatore.

Abstract
The concept of the customer segmentation is to target and done using the customer segmentation process using the clustering technique. Customer Segmentation using K-Means Algorithm" is based on the Analyzing such data is an important need. In the modern era of innovation, where there is a large competition to be better then everyone, the business strategy needs to be according to the modern conditions. The business done today runs on the basis of innovative ideas as there are large number of potential customers who are confounded to what to buy and what not to buy. The companies doing the business are also not able to diagnose the target potential customers. Here, the clustering algorithm used is K-means algorithm which is the partitioning algorithm, to segment the customers according to the similar characteristics. Customer Segmentation is the best application of unsupervised learning. Using clustering, identify segments of customers in the dataset to target the potential user base. They divide day which requires more work and time to do. To determine the optimal clusters, elbow method is used. Analyzing such data is an important need. In the modern era of innovation, where there is a large competition to be better then everyone, the business strategy needs to be according to the modern conditions. The business done today runs on the basis of innovative ideas as there are large number of potential customers who are confounded to what to buy and what not to buy. The companies doing the business are also not able to diagnose the target potential customers. The concept of which customer segment to target is done using the customer segmentation process using the clustering technique. The clustering algorithm used is K-means algorithm which is the partitioning algorithm, to segment the customers according to the similar characteristics. A research design is a matter plan specifying the methods and procedures for collecting and analyzing the needed data. Research design based on the exploratory research technique method and analysis is made on primary data collection for the project study. To determine the optimal clusters, elbow method is used.

INTRODUCTION

Nowadays the competition is vast and lot of technologies came into account for effective growth and revenue generation. For every business the most important component is data. With the help of grouped or ungrouped data, we can perform some operations to find customer interests. Internet users can choose from various online platforms to browse, compare, and purchase the items or services they need. While some websites specifically target B2B (business-to-business) clients, individual consumers are also presented with a vast number of digital possibilities (Muhammad Sufyan ,. Muhammad Sufyan and Deepak Srinivastava, 2022)Leading the global ranking of online retail websites in terms of traffic: The Seattle-based e-commerce giant that offers e-retail, computing services, consumer electronics, and digital content registered over 5.2 billion unique visitors in June 2020.

Data mining helpful to extract data from the database in a human readable format. But, we may not known the actual beneficiaries in the whole dataset**(**Jiawei Han, MichelineKamber and Jian Pei**,**2019**)**.Customer Segmentation is useful to divide the large data from dataset into several groups based on their age, demographics, spent, income, gender, etc. These groups are also known as clusters.

By this, we can get to know that, which product got huge number of sales and which age group are purchasing etc. And, we can supply that product much for better revenue generation.
Initially we are going to take the old data. As we know that old is gold so, by using the old data we are going to apply K-means clustering algorithm and we have to find the number of clusters first. So, at lastly, we have to visualize the data(T. Kanungo, D. M. Mount & N. S. Netanyahu**,**2002**)**. One can easily find the potential group of data while observing that visualization. The goal is to identify customer segments using the data mining approach, using the partitioning algorithm called as K-means clustering algorithm. The elbow method determines the optimal clusters.

REVIEW OF THE LITERATURE:
Customer Segmentation using K-Means Algorithm is based on the Analysing such data is an important need. In the modern era of innovation, where there is a large competition to be better then everyone, the business strategy needs to be according to the modern conditions**(YashKushwaha, Deepak kumarSrinivastva ,2020)**. The business done today runs on the basis of innovative ideas as there are large number of potential customers who are

confounded to what to buy and what not to buy. Concept decompositions for large sparse text data using clustering, is to analyse high dimensionality of text can be a deterrent in applying complex learners such as Support Vector Machines to the task of text classification**(S. Dhillon and D. M. Modha ,2001)**. Feature clustering is a powerful alternative to feature selection for reducing the dimensionality of text data. Here, we propose a new information theoretic divisive algorithm for feature/word clustering and apply it to text classification. An efficient K-means clustering algorithm, In k-means clustering, given a set of n data points in d-dimensional space Rd and an integer k and the problem is to determine a set of k points in Rd, called centers, so as to minimize the mean squared distance from each data point to its nearest center (**T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu,2002**). A popular heuristic for k-means clustering is Lloyd's algorithm present in a simple and efficient implementation of Lloyd's k-means clustering algorithm, which we call the filtering algorithm. The Basis Of Market Segmentation, proposes the effective decisions are mandatory for any company to generate good revenue. In these days competition is huge and all companies are moving forward with their own different strategies. We should use data and take a proper decision. Every person is different from one another and we don't know what he/she buys or what their likes are (**D. Aloise, A. Deshpande, P. Hansen, and P. Popat,2009**). But, with the help of machine learning technique one can sort out the data and can find the target group by applying several algorithms to the dataset.

STATEMENT OF PROBLEM
Customer Segmentation is the best application of unsupervised learning. Using clustering, identify segments of customers in the dataset to target the potential userbase. They divide day which requires more paperwork and time to do. As new technologies were emerging in today'sworld. Machine Learning which is powerful innovation which is used topredict the final outcome which has many algorithms. So for our problem statement we will use K-Means Clustering which groups the data into different clusters based on their similar characteristics and then we will visualize the data.

OBJECTIVE OF THE STUDY
To Segment the online shoppers and identify the different clusters based K-means clustering.

SCOPE OF THE STUDY
The scope of the study is to perform the users' Customer segmentation of shopping websites based on online buying behavior. First of all, it describes the research background and literature

review on sentimental polarity analysis of online buying behavior and its value mining; next it explains the research process of sentimental polarity analysis and how to process available data, preprocess text content, extract the main attributes of commodity or service, and identify the relationship between attributes and emotional words; and then it provides the reference for consumers, commodity producers and online shopping platforms according to analysis results; at the end it comes up with the foresight and future work.

METHODOLOGY

This paper aims to examine the sentimental analysis of a customer segmentation of an customer in an Retail Industry. I started to search for recent article (past five years form 2021) in ProQuest and Research Gate using a combination of terms related to sentimental analysis, Clustering used for segmentation and Customer segmentation based on buying behavior.

RESEARCH DESIGN: A research design is a matter plan specifying the methods and procedures for collecting and analyzing the needed data. Research design based on the exploratory research technique method and analysis is made on primary data collection for the project study.

SAMPLING FRAMEWORK:

Population:

Since,the exact number of customers shopping in online was unable to be found from the popular online shopping websites such as Amazon,Flipkart,Snapdeal. Hence,it was unknown population.

Sampling Type:

In this study, the convenience sampling type was followed where the reality data was collected from the customer in Online Retail shopping.

Sample size:

Since K-means algorithm are used for sample size more than 1500 would be more comfortable and hence got a sample of 3330.

DATA COLLECTION:

The data was collected from Online Retail Shoppers in structured questionnaire through Google form.

ANALYSIS USED:

The analysis used for customer segmentation for online shoppers is,

- K-means clustering
- Using Python Colab

ANALYSIS AND INTERPRETATION

SOURCE CODE:

```
import pandas as pd
df = pd.read_csv('consumerresponses.csv')
df
```

	Gender	Age	Academic Qualification	Products purchased in past year(Multiple answers)	Money spent on these purchases (Yearly)	Which website(s) do you frequently visit? (Multiple answers)	I make a purchase from website(s) on the basis of its design and graphics (colors,layout and images etc.).	I only make a purchase from website(s) if it displays branded products.	I feel comfortable to purchase from a website with interactive features.	I study terms and conditions before making an online payment.	...	What level of emotion do u see? Rejoice and love	What level of emotion do u see? Aggression and Relaxation	What le of emot do u s Thought and Str
0	Female	23	Doctrate (even if pursuing)	All products	10000	Snapdeal, Flipkart, Amazon	5	5	5	5	...	5	5	
1	Male	24	Post Graduate	Electronic goods(Mobiles etc)	10500	Snapdeal, Flipkart, Amazon	4	5	5	2	...	3	4	
2	Female	17	Post Graduate	Electronic goods(Mobiles etc), Clothing/Shoes/	1000	Snapdeal, Flipkart, Amazon	4	3	4	2	...	5	3	

Fig No 8: Importing data set file to the pandas library

INTERPRETATION:

First, import pandas library to analyze the data in the Python Colab. Get Data frame of your CSV file to read the dataset that you have collected from the customers. Print data frame to display entire data of your dataset.

DESCRIPTIVE ANALYSIS

SOURCE CODE:

```
import matplotlib.pyplot as plt
import seaborn as sns
genders = df.Gender.value_counts()
sns.set_style("darkgrid")
plt.figure(figsize=(12,5))
sns.barplot(x=genders.index, y=genders.values)
plt.show()
```

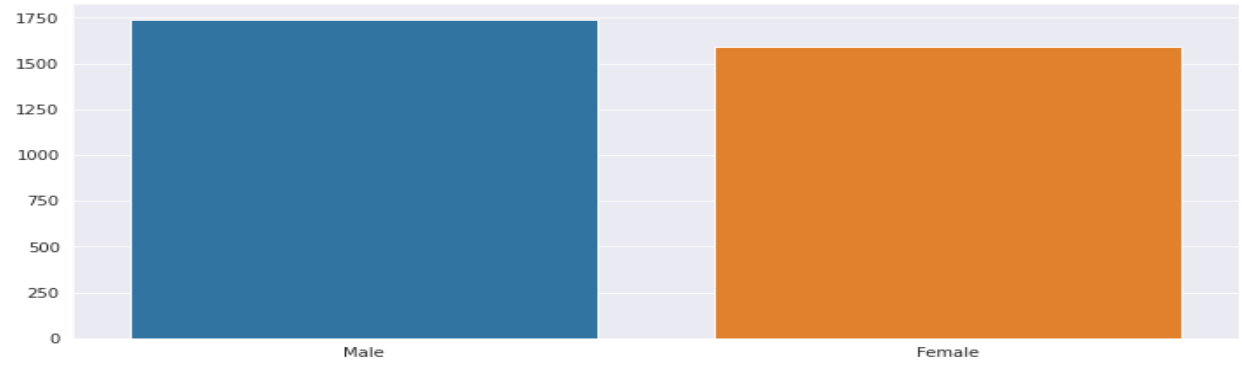

Chart No 1: Using Bar plot to display gender comparison

INTERPRETATION:

The bar plot to check the distribution of male and female population in the dataset. The male population 54% that clearly outweighs the female population of 46% of counterpart.

SOURCE CODE:

```
age18_25 = df.Age[(df.Age <= 25) & (df.Age >= 18)]
age26_35 = df.Age[(df.Age <= 35) & (df.Age >= 26)]
age36_45 = df.Age[(df.Age <= 45) & (df.Age >= 36)]
age46_55 = df.Age[(df.Age <= 55) & (df.Age >= 46)]
age55above = df.Age[df.Age >= 56]
x = ["18-25","26-35","36-45","46-55","55+"]
y = [len(age18_25.values),len(age26_35.values),len(age36_45.values),len(age46_55.values),len(age55above.values)]
plt.figure(figsize=(15,6))
sns.barplot(x=x, y=y, palette="rocket")
plt.title("Number of Customer and Ages")
plt.xlabel("Age")
plt.ylabel("Number of Customer")
plt.show()
```

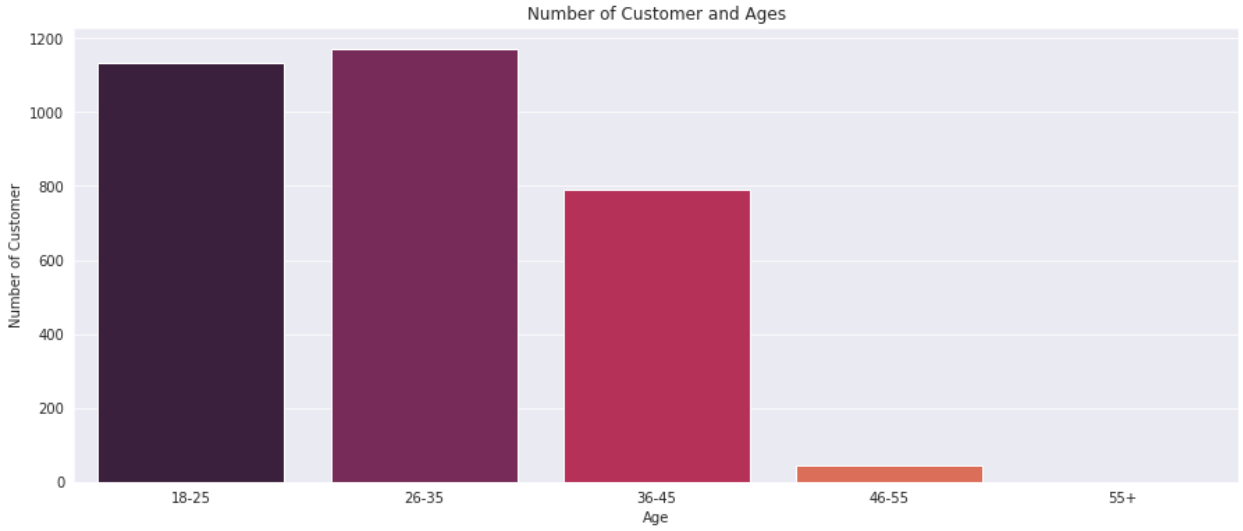

Chart No 2: Number of customers based on different age groups

INTERPRETATION:

Bar plot to check the distribution of number of customers in each age group. The age group between 26–35 agethat outweighs 35% every other age group. Whereas, 46-55 age group of customers that shows only 3% of customers who visit online website rarely.

USER CLASSIFICATION RESULTS

SOURCE CODE:

```
pca = PCA(n_components=14)
pca.fit(df1)
pca_fit = pca.transform(df1)
```

```
pca_fit = pd.DataFrame(pca_fit,index=df1.index)
for n in [3,4,5,6,7,8]:
    km = KMeans(n_clusters=n,random_state=0)
    clusters = km.fit_predict(pca_fit)
    silhouette_avg = silhouette_score(pca_fit, clusters)
    print("For n_clusters =", n,
          "The average silhouette_score is :", silhouette_avg)
    silhouette_values = silhouette_samples(pca_fit, clusters)
```

```
For n_clusters = 3 The average silhouette_score is : 0.11187811017444213
For n_clusters = 4 The average silhouette_score is : 0.07323184263424755
For n_clusters = 5 The average silhouette_score is : 0.06586592196650613
For n_clusters = 6 The average silhouette_score is : 0.06865134263157581
For n_clusters = 7 The average silhouette_score is : 0.06716231864317586
For n_clusters = 8 The average silhouette_score is : 0.07665563258444662
Kmeans_final = KMeans(n_clusters=3,random_state=0).fit(pca_fit)
df['cluster'] = Kmeans_final.labels_
fig = plt.figure()
ax = Axes3D(fig)
ax.scatter(pca_fit[0], pca_fit[2], pca_fit[1],c=df['cluster'],cmap=cm.hot)
plt.title('Data points in 3D PCA axis', fontsize=20)
plt.show()
```

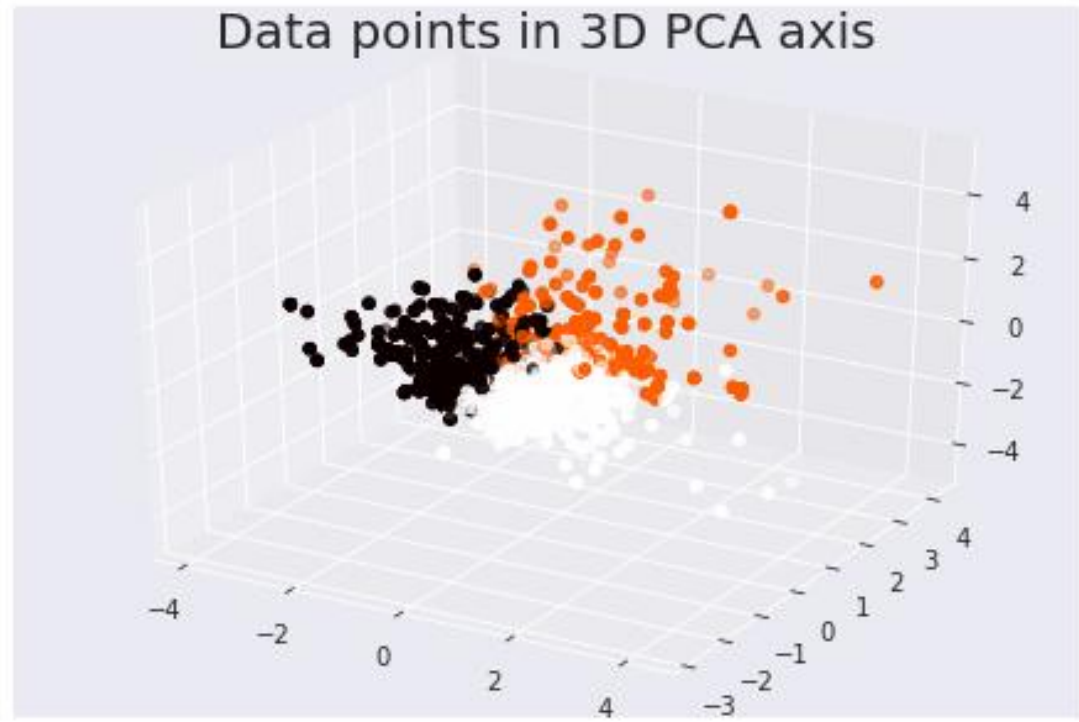

Chart No 3: K-means clustering using Silhouette values of the variables in the dataset

INTERPRETATION:

The k-means cluster is used to cluster the data. Judging by the elbow method , the decrease in Sum of Square Error is not significant when K is higher than 4. Hence, choosing K = 4 would yield favorable result. A Python colab library in Python language is used to implement the K-means algorithm, and the results are shown in Chart no: 3.

In the plot, X axis represents the Online purchase helps you to conserve time and money, Y axis represents the Money spent on these purchases (Yearly) and Z axis represents What level of emotions do you see? Happy and Giggly. It can be seen that the overall user data are close to 0 on X axis that represents the time and money. In the range method, the customer with the highest total purchase amount is taken as the maximum value. The plot shows that a small number of customers far exceed the average purchase amount.

It can be seen from the above customer purchase data that these customers have a larger total purchase amount, a higher purchase frequency, and a shorter time since last purchase, and the combined result of which is better performance on all three axes, thus making them appear as outliers of varying degrees.

At the same time, the indicators of different groups of customers and those of all customers as a whole are also extracted for analysis. The three outlier of money spend on these purchases extracted previously can be traced to Group 2, which is characterized by a large total purchase amount and a high purchase frequency. These users' purchase data are therefore in line with the overall characteristics of this group, thus proving rationality of the above clustering. Comparing the customer indicators of each group with the averages of those of all customers leads to the following findings.

Comparing the customer indicators of each group with the averages of those of all customers leads to the following findings.

Customers in Group 1 have a longer time since last purchase. The purchase data they left on the platform are less noticeable due to earlier time of last purchase, smaller value of total purchase amount, and lower purchase frequency. Moreover, the number of such customers is relatively small. /ey can be regarded as customers with loss risks and requiring further observation. Certain resources should be invested on the platform to further analyze and understand such customers.

The purchase frequency and total purchase amount of customers in Group 2 are greater than overall averages, and their last purchase is also more recent, indicating that online purchase converse time and money of customers. By bringing higher cash flow and profit to the platform and time consumption, they constitute a group of high-value customers. The platform should put more effort into maintaining and improving the relationship with them.

The total purchase amount and purchase frequency of Group 3 customers are more, and they completed their last purchase at level of emotions. This implies that inspite of their recent purchase behavior on the platform, they have formed a consumption habit there in a position to generate great level of emotions. They can be viewed as typical customers.

ELBOW METHOD:

SOURCE CODE:

```
wcss = []
for k in range(1,5):
   kmeans = KMeans(n_clusters=k, init="k-means++")
   kmeans.fit(df.iloc[:,6:20])
   wcss.append(kmeans.inertia_)
plt.figure(figsize=(12,6))
plt.grid()
plt.plot(range(1,5),wcss, linewidth=2, color="red", marker ="8")
```

```
plt.xlabel("K Value")
plt.xticks(np.arange(1,5,1))
plt.ylabel("WCSS")
plt.show()
```

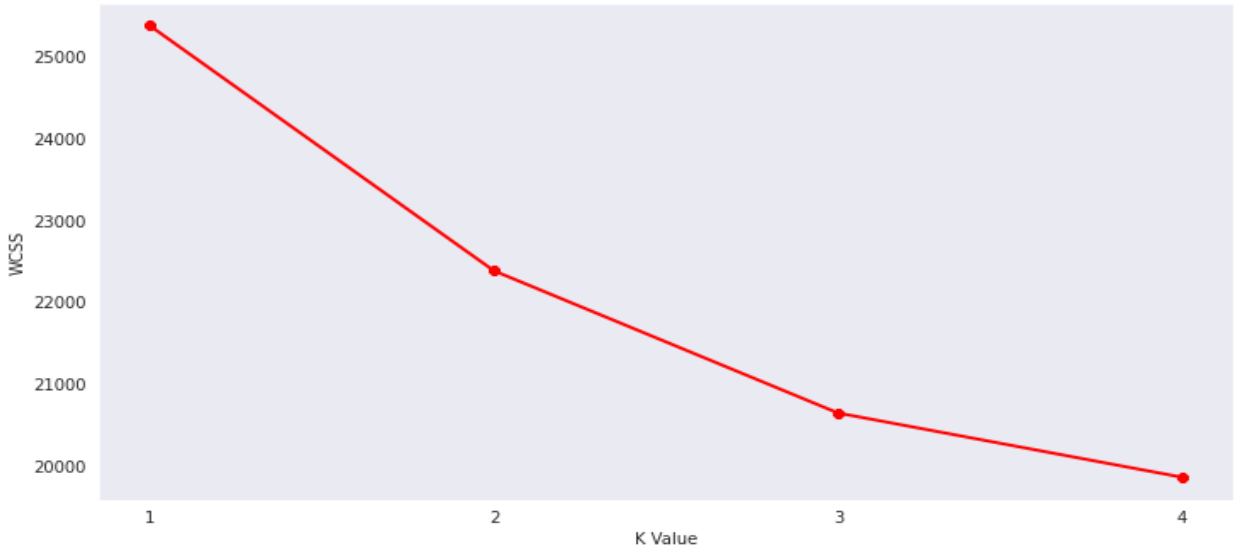

Chart No 4: Elbow method for two variables age and spending score of the customer

INTERPRETATION:

The elbow method is based on the observation that increasing the number of clusters can help to reduce the sum of within-cluster variance of each cluster. This is because having more clusters allows one to capture finer groups of data objects that are more similar to each other. To define the optimal clusters, Firstly, we use the clustering algorithm for various values of k. This is done by ranging k from 10 to 15 clusters. Then we calculate the total intra-cluster sum of square. Then, we proceed to plot intra-cluster sum of square based on the number of clusters. The plot denotes the approximate number of clusters required in our model. Here, the optimal number of cluster is 4.

CONCLUSION

In this research, a new and comprehensive approach to the study of online consumer behavior was introduced that explored this complex phenomenon from different angles. A broader picture was provided by understanding consumer macro-behavior. It illustrated the degree, purpose and attributes of Internet usage in a market and across multiple retailers. The intensive analysis of individuals indicated the detailed behavioral patterns and the reason for their variations. It depicted the complexity of online purchase decision-making processes and provided evidence for their dependence on individual differences and market characteristics.

Individual analysis proposed three segments of online consumers, based on the two individual characteristics of decision-making style and knowledge of the product. Combination of these two characteristics made it possible to describe behavioral variations, and has theoretical and practical implications. The attributes of behavior assigned to each segment can be constant or depend on the market

However, behavior in relation to intensity of decision-making cycles, duration of the process and the process outcome is a function of individual characteristics as well as the characteristics of the sector, depending on its importance and frequency of purchase. Specific measures are required to assist online purchase decision-making processes based on the needs of each segment of consumers. Several

recommendations were suggested. Finally, the results suggested a multi-channel strategy for current retailers, now and in the near future, contradicting the early perception of the Internet as a replacement channel.

REFERENCES

[1] MuhammadSufyan ,Deepak KumarSrivastava, “Understanding the customers’ shopping experience on online buying”,”Journel of retailing and consumer service”,vol.68,pp.111-123,2022
[2] I. S. Dhillon and D. M. Modha, “Concept decompositions for large sparse text data using clustering,” Machine Learning, vol. 42, issue 1, pp. 143-175, 2001.
[3] T. Kanungo, D. M. Mount, N. S. Netanyahu, “An efficient K-means clustering algorithm,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, pp. 881-892, 2002.
[4] MacKay and David, “An Example Inference Task: Clustering,” Information Theory, Inference and Learning Algorithms, Cambridge University Press, pp. 284-292, 2003.
[5] Jiawei Han, MichelineKamber, Jian Pei “Data Mining Concepts and Techniques”, Third Edition.
[6] D. Aloise, A. Deshpande, P. Hansen, and P. Popat, “The Basis Of Market Segmentation” Euclidean sum-of-squares clustering,” Machine Learning, vol. 75, pp. 245-249, 2009.
[7] S. Dasgupta and Y. Freund, “Random Trees for Vector Quantization,” IEEE Trans. on Information Theory, vol. 55, pp. 3229-3242, 2009.
[8]Puwanenthiren Premkanth, —Market Segmentation and Its Impact on Customer Satisfaction with Especial Reference to Commercial Bank of Ceylon PLC.‖ Global Journal of Management and Business Research Publisher: Global Journals Inc. (USA). 2012. Print ISSN: 0975-5853. Volume 12 Issue 1.
[9] Cooil, B., Aksoy, L. &Keiningham, T. L. (2008), ‘Approaches to customer segmentation’, Journal of Relationship Marketing 6(3-4), 9–39.
[10] D. Aloise, A. Deshpande, P. Hansen, and P. Popat, “The Basis Of Market Segmentation” Euclidean sum-of-squares clustering,” Machine Learning, vol. 75, pp. 245-249, 2009.
[11] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R.Silverman, and A. Y. Wu, “An efficient K-means clustering algorithm,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, pp. 881-892, 2002.
[12] Bhatnagar, Amit,Ghose, S. (2004), ‘A latent class segmentation analysis of e-shoppers’, Journal of Business Research 57, 758–767.
[13] Marcus, C. (1998), ‘A practical yet meaningful approach to customer segmentation approach to customer segmentation’, Journal of Consumer Marketing15, 494–504.