CIT. Journal of Computing and Information Technology, Vol. 26, No. 1, March 2018, 57–68
doi: 10.20532/cit.2018.1003863 

# Market Segmentation Analysis and Visualization Using K-Mode Clustering Algorithm for E-Commerce Business

Deepali Kamthania, Ashish Pahwa and Srijit S. Madhavan

School of Information Technology, VIPS, GGSIP University, Delhi, India

Now, all business organizations are adopting data-driven strategies to generate more profits out of their business. Growing startups are investing a lot of funds in data economy to maximize profits of the business group by developing intelligent tools backed by machine learning and artificial intelligence. The nature of business intelligence (BI) tool depends on factors like business goals, size, model, technology, etc. In this paper, the architecture of BI tool and decision process has been discussed with a focus on market segmentation, based on user behavior geographical distributions. Principal Component Analysis (PCA) followed by k-mode clustering algorithm has been used for segmentation. The proposed toolkit also incorporates interactive visualizations and maps.

*ACM CCS (2012) Classification:* Information systems → Information systems applications → Data mining → Clustering

Computing methodologies → Machine learning → Machine learning approaches → Factorization methods → Principal component analysis

Applied computing → Electronic commerce → Online shopping

*Keywords:* clustering, market segmentation, data visualization, principal component analysis (PCA), k-mode

## 1. Introduction

In today's business market, segmentation is applied in most market analysis processes for formulation of business policies. In the industrial environment, diverse segments come together and compete for delivery of merchandise and services for consumers [1]. There is a huge potential for business opportunities and need of visualization for planning stratergies [2], [3] and [4]. In the present scenario, existing visualization tools simply act as measurement dashboards, static depictions of organizational networks and not as exploratory systems whereas the market demands interactive business tools. The circle segment technique using one color pixel per data value [5] has been considered as an expressive powerful tool for visualizing large amounts of high dimensional data [6]. To connect with potential customer market segmentation is one of the commonly used marketing strategies, which segregates the consumers with common need into groups/subsets (segments) and then target them with distinct marketing mix [7], [8] and [9]. Market segmentation facilitates market diversity for formulation of strategies to maximized segment profit margins [10], [11]. The customer segmentation improves profits through direct marketing, especially catalog, mailers etc [12]. Target marketing is a tailored approach to capture segment of customers with customized plans and schemes [13]. "The best predictor of future customer behavior is past customer behavior" [14]. For over 50 years RFM (*Recency*, *Frequency* and *Monetary* value of the purchase) analysis has been used for segmentation for future customers [15]. Market segmentation is done on the basis of segmentation, targeting and positioning against the competitors by dividing a heterogeneous market into small homogeneous markets [16]. It is difficult to achieve appropriate levels of competitive insight in business ecosystems, under or over estimation can cause serious business loss or profit [17]. In a changing environment segmentation helps in determination of potential market. With the exponential growth in the volume of data over the internet, it becomes difficult to predict customer segment and pattern using traditional mathematical models. Timely identification of newly emerging trends from the huge volume of data plays a major role in the business process and decision making. The huge volume of data exists, but companies starve for knowledge. Data mining techniques help to overcome this knowledge scarcity.

Data mining is "knowledge discovery in databases (KDD) process" [18] to extract previously unknown, interesting patterns as groups of related data records (cluster analysis), unusual records (anomaly detection), and dependencies (association rule and sequential pattern mining) for future predictions [17], [19] and [20] from customer data due to their relatively low computational requirements. Clustering is one of the commonly used market segmentation techniques [21]. Clustering techniques minimize intra cluster distance and maximize inter cluster distance for data segmentation. Clustering methods can be broadly classified into four categories (17–25): Partitioning, Hierarchical, Density-based and Grid-based methods. Clustering can be applied in information retrieval, web pages grouping, image and market segmentation etc. [24]. For segment identification, customer clustering is applied [25] using customer characteristics (demographics, socioeconomic factors, and geographic location), product-related behavioral characteristics (purchase behavior, consumption behavior, preference for attractions, experiences, and services). ANN and k-means clustering are widely used for market segmentation [26], [27], [28]. Clustering can also be applied for evaluating supermarket shopping paths [29] or deriving employers' branding strategies [30] apart from customer segmentation. In order to extend distribution channel of the high end US furniture market, a socio-demographic segmentation for identification of potential buyer clustering technique has been used for analysis [31]. Data mining clustering technique has been applied to identify high-profit, high-value and low-risk customers. The data cleansing and patterns identification have been performed through demographic clustering algorithm using IBM I-Miner and data profiling has been done to develop clusters to identify high-value low-risk customers. It has been observed that a cluster typically represented 10–20 percent of customers which yields 80% of the revenue [32]. The SOM network developed using Iranian telecommunication company customer sales data showed that not only market segments but also mutual relationships between all market variables play major role [33]. Self-Organizing Map (SOM) has also been applied in tea-beverage for market segmentation and has showed improved results [34]. K-mean clustering has been applied for bicycle market segmentation. The different sub-segments were identified and separate policies were framed as per the segment characteristics for increasing the bicycles sales [35]. Market researchers figured out that segments are based on realistic conditions whereas cluster analysis allows segmentation on subjectivity basis [9]. Customer/market segmentation is the process of finding homogenous sub-groups within a heterogeneous aggregate market in direct marketingin order to focus on profitable market segments [36]. To apply product marketing strategies, customer segmentation has been carried out on a beverage distribution firm using modified k-means algorithm, applying GRASP philosophy to get the initial centers. The meta-heuristic proved to be robust and fast compared to existing methods [37]. The Demand Side Platforms DSPs' strategy for market segmentation and a selection model of the granularity for segmenting (Real Time Bidding) RTB advertising markets have been studied and it has been observed that market segmentation has a crucial impact on the RTB advertising effect and DSPs should adjust their market segmentation strategies according to their total number of advertisers [38].

In this paper, an attempt has been made to propose a model for formulating business strategies based on the users' interest and location. The clustering technique has been applied to customer's product-click data for segmentation and PCA technique has been applied to reduce dimensionality. Further, the geographical location has been fetched from an e-commerce website for data visualization.

## 2. Proposed Architecture

The streaming approach is risky, mainly because of the following reasons:

- Process generating the stream changes over time.
- Once two clusters are joined, there is no way to split the clusters required by the changes in the stream at a later stage.

The offline approach gives us an advantage of studying changes in clusters over time by storing the clusters at regular intervals. Moreover, offline methods give marketers flexibility to pull any subset of data and analyze the clusters formed during particular time intervals, which might help them retrieve greater insights and change trend from the data without dealing with the complexity of streaming architecture. Therefore, this paper laid a focus on system architecture, the flow of data and practical implementation in the business context.

Figure 1 shows high level 5 phased model of the proposed design. The click data generated by the user is loaded into the system through web user interface and recorded in a database. All data is fetched from the database through a Restful HTTP API service and sent to Cluster Identification Module which is responsible for segmenting the users into k clusters. Data Visualization Module has been used to plot latitude and longitude values of user location on an interactive map.

The proposed architecture has been implemented using the following software. The module for Data Loading, Cleaning and Transformation has been developed using Python 3 Jupyter Notebook. The Django Framework and Apache Web Server have been used to develop REST Service to provide value for data visualization. SQLite Database has been used for data storage and RStudio has been used to simulate the PCA algorithm and Data Visualization.

## 3. Mobile/Web User Interface

This section discusses the implementation of web user interface. At the beginning of every new session, the user is asked to provide the location on the web client. For every item click, the tracker function is called which sends the item name corresponding to the current user session to the controller function which is then written to the database along with the timestamp. This is shown in Algorithm 1.

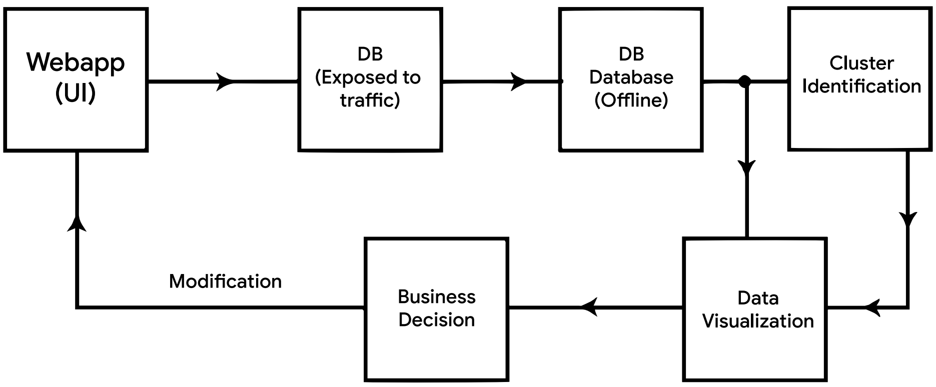


*Figure 1.* The architecture of proposed system.

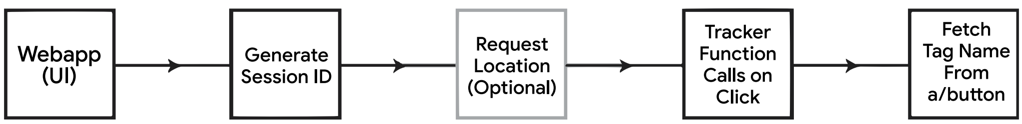


*Figure 2.* Web user interface process.

*Algorithm 1.* User Session Details.

1. Get *unique session* ID from 'data-username' attribute.
2. If unique session ID is not set:
   (i) Generate new unique session ID with random values.
   (ii) Store generated ID into the database.
3. Create new *temporary session* (with timestamp as the unique parameter) –
   (i) Store session in sessionStorage of the browser.
4. Get *latitude* and *longitude*.
5. Get *pathname* and *hostname*.
6. Insert all required session details to database: Current Session, Path Name, Hostname, Latitude, Longitude, Unique Session ID
7. ('a' tag or 'button' tag with data attribute – *data-analysis* = *TagName*)
8. Get tag name from the clicked attribute.
9. Update session details in the database: *TagName – where Temp Session is the same*.

## 4. Database

The dataset refers to 200 users of dummy e-commerce website developed and hosted online as a part of the experiment conducted for this research. It displayed 68 items for sale. Users were asked to click on the items in order of their preference. Every click for each current user session is registered in an online database which tracks the order in which the items are clicked by the user. At some time $t$, the data is pulled by a remote machine and loaded in the memory for further analysis, as shown in Figure 3.

## 5. User Segment Identification Module

In this module, collected data is transformed into $N \times n$-dimensional sparse Boolean feature matrix $M$, where each row represents a user session and columns represent items such that,

$M[i, j] = 1$, if user $i$ clicked on the item $j$,

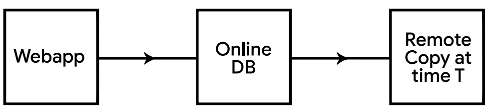


*Figure 3.* Record fetched from the server database.

$M[i, j] = 0$, otherwise

$N$ = Number of items

$n$ = Number of user sessions

PCA is performed on this matrix to identify the features in which most of the information is contained, thereby reducing the dimensionality of the original dataset.

### 5.1. Principal Component Analysis (PCA)

PCA maps the number of correlated variables into a smaller set of uncorrelated variables called the principal components. The first principal component represents maximum variability in the data and succeeding components describe the remaining variability [39]. PCA and the linear transformation are used for dimensionality and noise reduction and initial centeroid computation for k-means clustering algorithm [36] and [40]. The heuristics approach is used to reduce the number of distance calculation to assign the data point to the cluster. It has been analyzed that as the number of records increases, the time execution of both techniques is increased. It has also been observed that fuzzy c-means performs better than the k-means algorithm [41]. Fuzzy c-means, compared with respect to other Clustering Al-

*Algorithm 2.* Algorithm for fetching records for analysis.

1. Initialize API endpoint object
2. Make post request to the endpoint
3. Parse the data in JSON format
4. Compute the number of records
5. Create database connection object
6. For each record fetch the following values: ID, Username, pathname, Host_name, valueKey, Session_value, Latitude, Longitude, tagData, dateVal.

*Algorithm 3.* PCA Implementation.

1. Load the data
2. Identify predictors
3. Identify and remove zero variance columns from the dataset
4. The PCA is performed using R function prcomp (). It centers the variable so it has the mean value equal to zero. It also normalizes the variables to have standard deviation equal to 1.

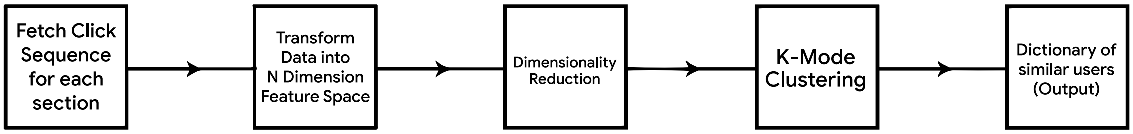


*Figure 4.* Flow of PCA and k-mode clustering implementation.

| | Armani | Polo | PE | Puma | Privecy Policy | twitter-ref | Bags | Guess | Acne | Company Information | ... | phone | Fashion | T-Shirt | Pantaloon Tee | fb-ref | Interiors | Versace | Polo |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **1** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **2** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **3** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **4** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |

*Figure 5.* Screenshot of dataset.

gorithm, requires more computation time [42]. In the current study, the purpose of the analysis of principal components is to compress information in a large set of correlated variables to few factors that can synthesize most of the total information contained in the original variables. The new set of features is sent as an input to k-mode clustering algorithm.

Figure 4 depicts the flow of data in the processing pipeline that transforms the raw data into clusters. Dimensionality reduction is performed using PCA followed by k-modes clustering algorithm to obtain a dictionary of similar users where the key represents cluster number and values are lists of similar users identified by session id.

Figure 7 shows that it was observed that 20 components explained approximately 98.29% of the variance. So, for k-mode, only those 20 components were taken into consideration.

```
> colnames(mydata)
 [1] "Adidas"                      "T.C"
 [3] "Video2"                      "Themeum.Designers"
 [5] "Nike"                        "Kids"
 [7] "Cotton.Fabrics"              "Guess"
 [9] "Easy.Polo.Black.Edition"     "Polo"
[11] "phone"                       "Under.Armour"
[13] "twitter.ref"                 "Designer.Tops"
[15] "Shoes"                       "Fendi"
[17] "PE"                          "Video1"
[19] "fb.ref"                      "fb.ref.1"
[21] "PE.1"                        "Easy.Polo.Black.Edition.1"
[23] "Valentino"                   "Clothing"
[25] "Nike.1"                      "Easy.Polo.Blue.Edition"
[27] "Guess.1"                     "Under.Armour.1"
[29] "Interiors"                   "Bags"
[31] "Rösch.creative.Clothing"     "Puma"
[33] "Versace"                     "Armani"
[35] "email"                       "Video1.1"
[37] "Video4"                      "Rösch.creative.Clothing.1"
[39] "Video4.1"                    "T.C.1"
[41] "phone.1"                     "Fashion"
[43] "T.Shirt"                     "Fendi.1"
[45] "Designer.Tops.1"             "Privecy.Policy"
[47] "T.Shirt.1"                   "Puma.Tee"
[49] "Pantaloon.Tee"               "Clothing.1"
[51] "Video3"                      "Company.Information"
[53] "linkedin.ref"                "Valentino.1"
[55] "Polo.1"                      "Video3.1"
[57] "Video2.1"                    "ASICS"
[59] "blog.post.1"                 "Versace.1"
[61] "Acne"                        "Company.Information.1"
[63] "dribble.ref"                 "gPuss.ref"
[65] "Designer.Blazers"            "Cotton.Fabrics.1"
[67] "Chanel"                      "Nike.2"
```

*Figure 6.* Snapshot of 68 boolean predictor variables.

## 6. Data Mining and Clustering Methods

The k-means [43] is the most commonly used clustering technique where the accuracy depends upon the choice of the initial seeds [44]

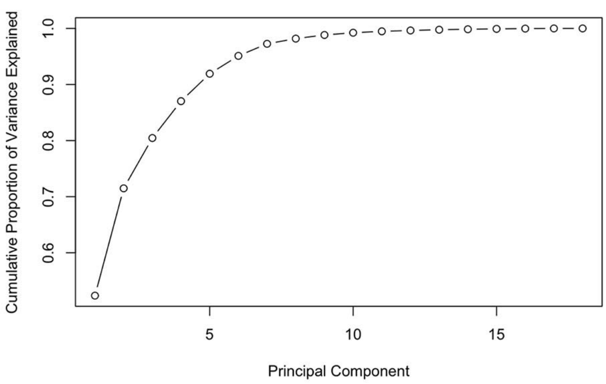


*Figure 7.* Principal component vs cumulative proportion of variance.

and it often falls in local optima, which is a major disadvantage for post-hoc market segmentation cases as accurate market clusters designation is not possible for market managers. Therefore, in the current study, k-mode clustering algorithm has been considered.

## 6.1. K-Mode Clustering Algorithm

K-mode clustering algorithm is an extension to the standard k-means clustering algorithm. Unlike k-means, k-mode operates on categorical data. The major modification includes distance function, cluster center representation and iterative clustering process [45] and [46].

### 6.1.1. The K-Modes Algorithm

Let $D = \{X_1, X_2, \ldots, X_n\}$ be a set of $n$ categorical objects, where each $X_i = [x_i,1, x_i,2, \ldots, x_i,m]$ is described by $m$ categorical attributes $A_1, \ldots, A_m$. Each attribute $A_j$ describes a domain of values, denoted by

$$Dom\left(A_j\right) = \left\{a_j^{(1)}, a_j^{(2)}, \ldots, a_j^{(p_j)}\right\}$$

where $p_j$ is the number of category values of attribute $A_j$. The k-modes algorithm operates on the k-means model to search a partition of $D$ into $k$ clusters such that it minimizes the objective function $P$ with unknown variables $U$ and $Z$ according to following equations: [47]

$$P(U,Z) = \sum_{l=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{m} u_{i,j} d\left(x_{i,j}, z_{i,j}\right) \tag{1}$$

Subject to

$$\sum_{l=1}^{k} u_{i,j} = 1, \qquad 1 \le i \le n \tag{2}$$

$$u_{i,j} \in \{0,1\}, \qquad 1 \le i \le n, \quad 1 \le l \le k \tag{3}$$

where

- $U$ is a $n \times k$ partition matrix, $u_{(i,j)}$, is a binary variable, and $u_{(i,j)} = 1$, indicates that object $x_i$ is allocated to cluster $C_i$;
- $Z = \{Z_1, Z_2, \ldots, Z_k\}$ is a set of vectors representing the centers of the $k$ clusters, where

$$Z_i = [Z_{l,1}, Z_{l,2}, \ldots, Z_{l,m}\} \quad (1 \le l \le k)$$

### 6.1.2. Distance Function

$d(x_{i,j}, z_{i,j})$ is a distance or dissimilarity measure between object $X_i$ and the center of cluster $C_i$ on attribute $A_j$. In k-modes algorithm, a simple matching distance measure is used. That is, the distance between two distinct categorical values is 1, while the distance between two identical categorical values is 0. More precisely [47],

$$d\left(x_{i,j}, z_{i,j}\right) = f(x) = \begin{cases} 0, & \left(x_{i,j} = z_{i,j}\right) \\ 1, & \left(x_{i,j} \ne z_{i,j}\right) \end{cases} \tag{4}$$

The optimization problem in k-modes clustering can be solved by iteratively solving the following two minimization problems [47]:

1. Problem $P_1$: $F_{ix}\, Z = \hat{Z}$, solve the reduced problem $P(U, \hat{Z})$
2. Problem $P_2$: $F_{ix}\, U = \hat{U}$, solve the reduced problem $P(\hat{U}, Z)$

Problems $P_1$ and $P_2$ are solved according to the two following theorems, respectively.

**Theorem 1:** Let $Z = (\hat{Z})$ be fixed, $P(U, \hat{Z})$ is minimized if and only if

$$u_{ij} = f(x) = \\ = \begin{cases} 1, & \sum_{j=1}^{m} d\left(x_{i,j}, z_{i,j}\right) \le \sum_{j=1}^{m} d\left(x_{i,j}, z_{h,j}\right) \ \forall h,\ 1 \le h \le k \\ 0, & \text{otherwise} \end{cases}$$

**Theorem 2:** Let $U = \hat{U}$ be fixed $P(U, \hat{Z})$ is minimized if and only if

$$z_{i,j} = a_j^{(r)}$$

where $a_j^{(r)}$ is the mode of attribute values of $A_i$ in the cluster $C_i$ that satisfies

$$f\left(a_j^{(r)} \middle| C_l\right) \ge f\left(a_j^{(t)} \middle| C_l\right) \ \forall t,\ 1 \le t \le p_j$$

Here $f\left(a_j^{(r)} \middle| C_l\right)$ denotes the frequency count of attribute value $a_j^{(r)}$ in the cluster $C_i$ i.e.

$$f\left(a_j^{(r)} \middle| C_l\right) = \left|\left\{u_{i,j} \middle| x_{i,j} = a_j^{(r)}, u_{i,j} = 1\right\}\right|$$

#### 6.1.3. The K-Modes Algorithm Implementation

1. Randomly choose an initial $Z^{(1)}$. Determine $U^{(1)}$ such that $P(U, Z^{(1)})$ is minimized. Set $t = 1$.
2. Determine $Z^{(t+1)}$ such that $P(U^{(t)}, Z^{(t+1)})$ is minimized. If $P(U^{(t)}, Z^{(t+1)}) = P(U^{(t)}, Z^{(t)})$, then stop; otherwise go to step 3
3. Determine $U^{(t+1)}$ such that $P(U^{(t+1)}, Z^{(t+1)})$ is minimized. If $P(U^{(t+1)}, Z^{(t+1)}) = P(U^{(t)}, Z^{(t+1)})$, then stop; otherwise $t = t + 1$ go to step 2

### 6.2. Cluster Analysis and Identification Process

To study the hidden underlying user clusters in the dataset, the assumed number of clusters (k) were iteratively incremented from 2 to 60 and the model was trained using the k-mode algorithm.

*Algorithm 4.* Evaluation Procedure.

**For** *i* in range (1, 60)
  Set number_of_clusters = *i*
  Train the model using the k-mode algorithm
  Obtain cluster centroids
**Exit**

#### 6.2.1. Cluster Evaluation

In the current study, clusters have been evaluated using Silhouette Score Evaluation Metric. Silhouette Score is a measure of similarity which measures the proximity of an object to its own cluster (cohesion) compared to other clusters (separation) [48]. The performance of clusters obtained at each iteration using the algorithm discussed in the previous section has been evaluated using Silhouette Score for the different number of pre-initialized cluster centers. The maximum value of Silhouette Score observed during the experiment was 0.967742, when the assumed number of underlying clusters was set to 31, as shown in Figure 8.

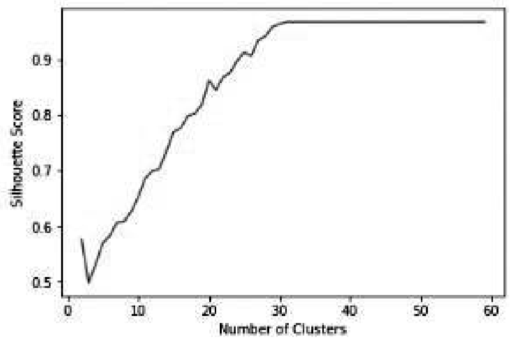


*Figure 8.* Number of clusters vs silhouette score.

For behavior models it is important to make some intuitive sense to humans. In this section, results of qualitative analysis of a few behavioral clusters are discussed. The e-commerce website used for the experiment consisted of 68 male and female garment categories. Each click triggered an event which was recorded as either 1 or 0 for that user session. It was observed that 68% of sessions that belonged to cluster #1 had a high preference for female category items. 57% sessions belonging to this cluster showed interest in mostly all Nike items presented on the site. In Cluster 3, most visitors did not click any item at all.

## 7. Mapping Geographical Distribution

Knowledge of the geographical distribution of consumer base helps business organizations to make better marketing strategies in order to improve their profits. This subsystem in the proposed model is designed to interactively study the geographical distribution of customers visiting e-commerce websites.

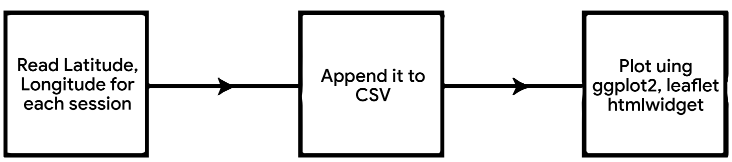


*Figure 9.* Mapping geographical distribution.

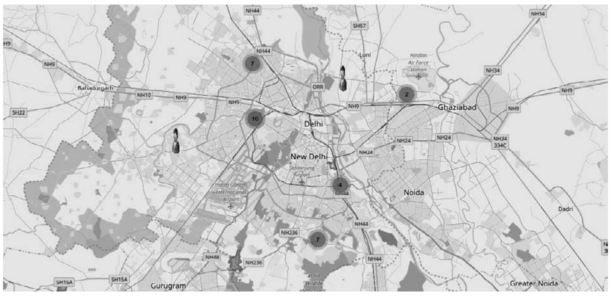


*Figure 10.* Geographical distribution of users generated by ggplot2 and leaflet.

Data is read from sqlite3 database and appended to csv file. It is then pulled by the program written in R which generates a map.html file with all the locations marked as different color points. The file is served on the browser through apache web server on a local system. The geocoordinates from the data collected during the visitor session are appended to a .csv file which is then plotted on an interactive map using ggplot and leaflet packages of R programming language.

## 8. Visualization

Visualizations helps business decision makers to examine a large amount of data quickly hiding the complexities of what is actually stored in databases or excel sheets. Thus the importance of data visualization in business intelligence cannot be ignored. The proposed subsystem aims to provide interactive time series, bar plots and graphs to answer some important business questions. The goal of the module is to answer the following important questions:

- Which was the most popular product on any given day?
- How the popularity of any product varies over time?

The architecture of the system is discussed in the following subsection.

Figure 11 shows the high-level architecture of visualization module developed on Django framework. It computes the count of clicks on each product for a range of days in order to visualize how a product is performing over a period of time. The counts are sent as json object for each product whenever a web client requests them and graphical representation is displayed on the browser, transformed to list of coordinates and pushed into the server. The information is fetched by a web application from the server which plots the information on the system of the decision maker, once it is available. A decision is made and the suggested changes are reflected back to the original e-commerce content displayed to users.

In Figure 12, $x$-axis shows name of product and $y$-axis shows total number of clicks on a selected day.

In Figure 13, $x$-axis shows date and $y$-axis shows number of clicks for a selected product.

## 9. Conclusions

The paper describes the BI tool and a decision process for market segmentation based on user behavior analysis and geographical information. PCA, followed by the application of the k-mode clustering algorithm, has been applied for segmentation. The proposed architecture

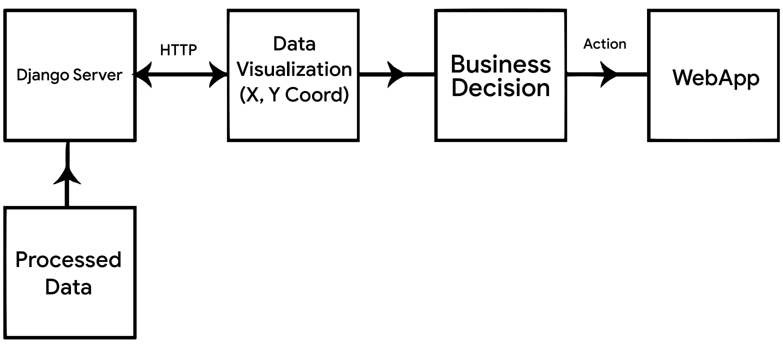


*Figure 11.* Architecture for data visualization.

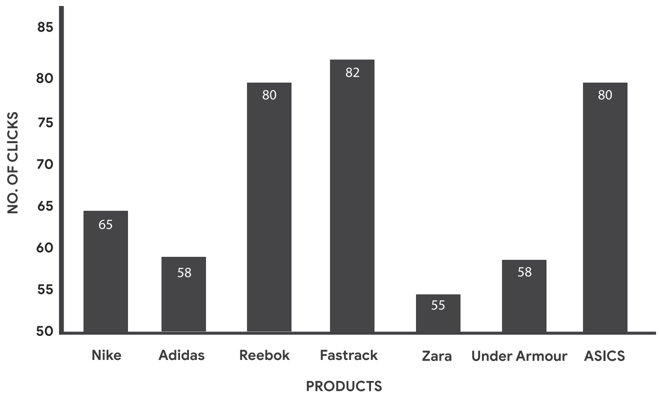


*Figure 12.* Number of clicks for products.

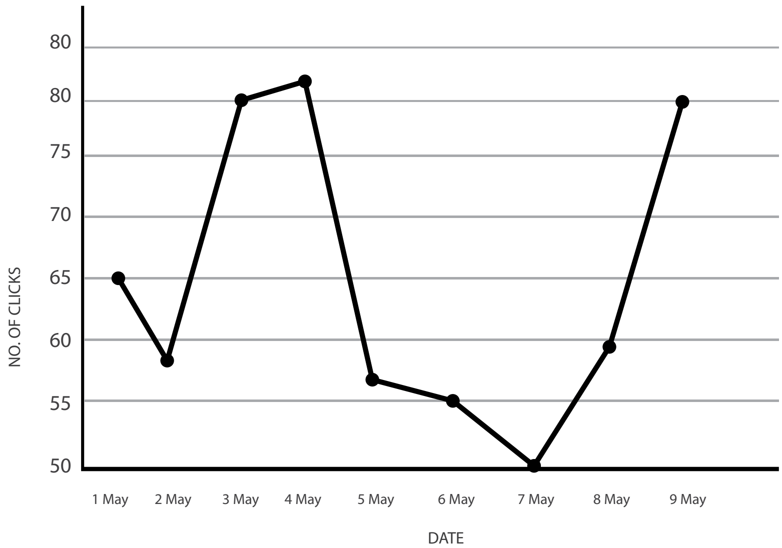


*Figure 13.* Day wise click of individual product.

provides a simplified system for formulating business strategies suitable for the small internet business owners or growing startups. It is a complete toolkit from data cleaning to visualization. The proposed system also identifies the popularity of each product within a time period using interactive visualizations in targeting the unique segments for connecting with potential customers for business expansion.

## 10. Future Work

The proposed architecture can be extended to integrate with real-time analytical tools. The proposed system can be scaled up by adopting HDFS and MapReduce techniques to build a fully fledged production system.

## References

[1] R. C. Basole and W. Rouse, "Complexity of Service Value Networks: Conceptualization and Empirical Investigation", *IBM Systems Journal*, vol. 47, no. 1, pp. 53–70, 2008. http://dx.doi.org/10.1147/sj.471.0053

[2] H. Chen *et al.*, "Business Intelligence and Analytics: from Big Data to Big Impact", *MIS Quarterly (Special Issue Business Intelligence Research)*, vol. 36, no. 4, pp. 1165–1188, 2012. https://ai.arizona.edu/sites/ai/files/MIS611D/chen-bi-december-2012.pdf

[3] R. Lengler and M. J. Eppler, "Towards a Periodic Table of Visualization Methods for Management", in *IASTED Proceeding of the Conference on Graphics and Visualization in Engineering (GVE 2007)*, 2007, pp. 83–88.

[4] K. Zhang, "Using Visual Languages in Management", J*ournal of Visual Languages & Computing*, vol. 23, no. 6, pp. 340–343, 2012. http://dx.doi.org/10.1016/j.jvlc.2012.09.001

[5] M. Ankerst *et al*., "Circle Segments: A Technique for Visually Exploring Large Multidimensional Data Sets", in *Proc. Visualization '96*, *Hot Topic Session*, San Francisco, CA, 1996. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.68.1811&rep=rep1&type=pdf

[6] D. A. Keim, "Pixel-oriented Visualization Techniques for Exploring Very Large Databases", *Journal of Computational and Graphical Statistics*, vol. 5, pp. 58–77, 1996. https://pdfs.semanticscholar.org/ce1e/b9ed41232690a1ab0b6b7322cfdb10a385cc.pdf

[7] P. Kotler, "Marketing Management – Analysis, Planning and Control" (4th ed.), Prentice-Hall, 1980.

[8] M. J. Crotft, "Market Segmentation: A Step-by-step Guide to Profitable New Business", Routledge, Paperback, 1994.

[9] M. Sarstedt and E. Mooi, "A Concise Guide to Market Research", Springer Texts in Business and Economics, Springer-Verlag Berlin Heidelberg, pp. 273–324, 2014. http://dx.doi.org/10.1007/978-3-642-53965-7

[10] A. Weinstein, "Market Segmentation: Using Demographics, Psychographics and other Niche Marketing Techniques to Predict and Model Customer Behavior", Chicago: Probus Pub. Co., 1994

[11] V. Venugopal and W. Baets, "Neural Networks and Statistical Techniques in Marketing Research: A Conceptual Comparison", *Marketing Intelligence and Planning*, vol. 12, no. 7, pp. 30–38, 1994. http://dx.doi.org/10.1108/02634509410065555

[12] David Shepard Associates, "The New Direct Marketing: How to Implement a Profit-Driven Database Marketing Strategy", 3rd edition, 1998.

[13] P. Kotler, "Marketing Management", Englewood Cliffs, Prentice-Hall, 2001.

[14] C. Swearingen, "101 Powerful Marketing Strategies for Growing Your Business Now!", *SmallBiz Marketing Services*, pp. 24–27, 2009.

[15] C. K. Bhensdadia and Y. P. Kosta, "Discovering Active and Profitable Patterns with RFM (Recency, Frequency and Monetary) Sequential Pattern Mining – a Constraint based Approach", *International Journal of Information Technology and Knowledge Management*, vol. 4, no.1, pp. 27–32, 2011. https://pdfs.semanticscholar.org/6771/423d0ac5347555c5ead347374175d3cd7ea2.pdf

[16] A. L. Majurin, "Industrial Segmentation. A Review", *Memo-Stencil, Preliminära Forskningsrapporter från Företagsekonomiska*, 208, 2001, ISSN 0357-4571. http://web.abo.fi/fak/esf/fei/wrkpaper/ms208/memostencil208.pdf

[17] R. Adner, "The Wide Lens: A New Strategy for Innovation", *Penguin Group*, vol. 30, no. 1, pp. 277, 2012. https://doi.org/10.1111/j.1467-9310.2012.00697 xhttp://www.vikalpa.com/pdf/articles/2014/vikalpa-39-1-159.pdf

[18] U. Fayyad *et al*., "From Data Mining to Knowledge Discovery in Databases", AI Magazine, vol. 17 no. 3, pp. 37–54, 1996. https://doi.org/10.1609/aimag.v17i3.1230

[19] T. Harris, "Optimization Creates Lean Green Supply Chains", Data Mining Book, 2008.

[20] M. Hartely, "Using Data Mining to Predict Inventory Levels", Data Mining Book, 2005.

[21] M. Wedel and W. A. Kamakura, "Market Segmentation Conceptual and Methodological Foundations", Kluwer Academic Publishers, 2000.

[22] I. H. Witten and E. Frank, (2000). Data mining: Practical machine learning tools and techniques. Morgan Kaufmann series in data management systems.

[23] S. Hwang and J. C. Thill, "Using Fuzzy Clustering Methods for Delineating Urban Housing Submarkets", *Proceedings of the 15th International Symposium on Advances in Geographic Information Systems*, *[17] IDC*, 2013, *IDC Big Data and Business Analytics Forum 2013*, *in Leveraging Data for Agile Business*. http://dx.doi.org/10.1145/1341012.1341031

[24] D. T. Pham and A. A Afify, "Clustering Techniques and their Applications in Engineering", *Proceedings of the Institution of Mechanical Engineers*, *Part C: Journal of Mechanical Engineering Science*, 2006, pp. 103–119. http://dx.doi.org/10.1016/B978-008045157-2/50060-2

[25] G. Saarenvirta, "Mining Customer Data", *DB2 Magazine*, vol. 3, no. 3, pp. 10–20, 1998.

[26] C. Hung and C. Tsai, "Market Segmentation based on Hierarchical Self Organizing Map for Markets of Multimedia on Demand", *Expert Systems with Applications*, vol. 34, pp.780–787, 2008. http://dx.doi.org/10.1016/j.eswa.2006.10.012

[27] K. C. Gehr and S. Shim, "A Shopping Orientation Segmentation of French Consumers: Implications for Catalog Marketing", *Journal of Interactive Marketing*, vol. 12, no. 4, pp. 34–46, 1998. http://dx.doi.org/10.1002/(SICI)1520-6653(199823)12:4%3C34::AID-DIR4%3E3.3.CO;2-F

[28] R. J. Kuo *et al*., "Integration of Self-organizing Feature Maps and Genetic-Algorithm-Based Clustering Method for Market Segmentation", *Journal of Organizational Computing and Electronic Commerce*, vol. 14, no.1, pp. 43–60, 2004. http://dx.doi.org/10.1207/s15327744joce1401_3

[29] J. S. Larson *et al.*, "An Exploratory Look at Supermarket Shopping Paths", *International Journal of Research in Marketing*, vol. 22, no.4, pp. 395–414, 2005.
http://dx.doi.org/10.1016/j.ijresmar.2005.09.005

[30] L. Moroko, and M. D. Uncles, "Employer Branding and Market Segmentation", *Journal of Brand Management*, vol. 17, no. 3, pp. 181–196, 2009.
http://dx.doi.org/10.1057/bm.2009.10

[31] L. E. ThiThu Hoa *et al.*, "Decision Support System based on Socio-demographic Segmentation and Distribution Channel Analysis in the US Furniture Market", *International Conference on Industrial Engineering and Systems Management (IESM)*, 2009, Montreal, Canada.

[32] S. Rajagopal, "Customer Data Clustering using Data Mining Technique", *International Journal of Database Management Systems (IJDMS)*, vol. 3, no. 4, pp. 1–11, 2011.
http://dx.doi.org/10.5121/ijdms.2011.3401
https://arxiv.org/ftp/arxiv/papers/1112/1112.2663.pdf

[33] P. Hanafizadeh *et al.* "An Expert System for Perfume Selection using Artificial Neural Network", *Expert Systems with Applications*, vol. 37, no. 12, pp. 8879–8887, 2010.
http://dx.doi.org/10.1016/j.eswa.2010.06.008

[34] C. W. Hong, "Using the Taguchi Method for Effective Market Segmentation", *Expert Systems with Applications*, vol. 39, no. 5, pp. 5451–5459, 2012.
http://dx.doi.org/10.1016/j.eswa.2011.11.040

[35] Z. Li *et al.*, "Bicycle Commuting Market Analysis using Attitudinal Market Segmentation Approach", *Transportation Research Part A: Policy and Practice*, vol. 47, pp. 56–68, 2013.
http://dx.doi.org/10.1016/j.tra.2012.10.017

[36] A. Fahmida *et al.*, "Comparative Performance of Using PCA With K-Means and Fuzzy C-Means Clustering For Customer Segmentation", *International Journal of Scientific and Technology Research*, vol. 4, no.10, pp. 70–74, 2015.

[37] L. Diana *et al.*, "An Iterated Greedy Heuristic for a Market Segmentation Problem with Multiple Attributes", *European Journal of Operational Research*, vol. 261, no. 1, pp. 75–87, 2017.

[38] R. Qin *et al.*, "Exploring the Optimal Granularity for Market Segmentation in RTB Advertising via Computational Experiment Approach", *Electronic Commerce Research and Applications*, vol. 24, pp. 68–83, 2017.
http://dx.doi.org/10.1016/j.elerap.2017.07.001

[39] S. S. Deshmukh *et al.*, "Pose Variant Based Comparative Analysis of PCA and LDA", *Second International Conference on Emerging Trends in Engineering & Technology*, 2009.
http://dx.doi.org/10.1109/ICETET.2009.181

[40] S. Tajunisha, "Performance Analysis of K-means with Different Initialization Methods for High Dimensional Data", *International Journal of Artificial Intelligence and Applications*, vol. 1, no. 4, pp. 44–52, 2010.
http://dx.doi.org/10.5121/ijaia.2010.1404

[41] A. Sheshasayee and P. Sharmila, "Comparative Study of Fuzzy C-Means and K Means Algorithm for Requirements Clustering", *Indian Journal of Science and Technology*, vol. 7, no. 6, pp. 853–857, 2014.
http://www.indjst.org/index.php/indjst/article/viewFile/47757/41449

[42] S. Tejwant and M. Manish, "Performance Comparison of Fuzzy C-Means with Respect to Other Clustering Algorithms", *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 4, pp. 89–93, 2014.
https://pdfs.semanticscholar.org/3720/5c8fe390d36bde67a2e0f614d5ce8bba829b.pdf

[43] J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations", in L. M. L. Cam & J. Neyman (Eds.), *Proceeding of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 281–297, 1996, University of California Press.

[44] G. W. Milligan and M. C. Cooper, "An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms", *Psychometrika*, vol. 45, no. 3, pp. 159–179, 1980.
http://dx.doi.org/10.1007/BF02293907

[45] J. Z. Huang, "Clustering Categorical Data with K-Modes", The University of Hong Kong, Hong Kong, pp. 246–248, 2009, IGI Global.
http://dx.doi.org/10.4018/978-1-60566-010-3.ch040

[46] A. M. Gayathri "Enhanced Customer Relationship Management", *International Journal of Computer Science & Engineering Technology*, vol. 1, no. 4, pp. 163–167, 2011.

[47] Z. He *et al.*, "Attribute Value Weighting in K-Modes Clustering", Department of Computer Science and Engineering, Harbin Institute of Technology, Expert Systems with Applications, vol. 38, no. 12, pp. 42–47, 2011.
https://arxiv.org/pdf/cs/0701013

[48] P. J. Rousseeuw, "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis", *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1984.

*Received*: November 2017
*Revised*: May 2018
*Accepted*: June 2018

*Contact addresses*:
Deepali Kamthania
School of Information Technology
VIPS, GGSIP University
Delhi-110034
India
e-mail: deepali102@gmail.com

Ashish Pahwa
School of Information Technology
VIPS, GGSIP University
Delhi-110034
India
e-mail: ashishpahwa7@gmail.com

Srijit S. Madhavan
School of Information Technology
VIPS, GGSIP University
Delhi-110034
India
e-mail: srijitsm123@gmail.com

Deepali Kamthania received the BSc and MCA degrees from Aligarh Muslim University (AMU) Aligarh in 1996, and 1999, respectively, and the PhD degree from the Indian Institute of Technology (IIT) Delhi in 2012. She has more than 15 years of experience in academics and IT industry, and is presently working as Professor in the School of Information Technology, VIPS, Delhi. Her areas of research interest include artificial neural networks, solar thermal applications and data mining. She has published over 50 research papers in reputed international and national journals and has delivered many invited lectures at different academic forums. She is the recipient of the Academic Excellence Awards, CSI and IETE Awards for significant contribution to the IT field. Dr. Kamthania is a life time member of IEEE, CSI and ISTE.

Ashish Pahwa completed his senior secondary education at Bloom Public School, VasantKunj, New Delhi in the 2015, and is currently pursuing his BCA degree from Vivekananda Institute of Professional Studies. In 2017, he worked as an intern for a number of companies: SigmaWay LLC as a Data Science intern, Pronto-IT Labs as a Big Data intern and Kreatik as a Backend Developer intern. In 2018, he joined Novuse Internet Pvt Ltd as a Data Engineer intern. He is President for the academic year 2017-18 of ACE – The CSI Student Branch of VIPS.

Srijit S. Madhavan completed his senior secondary education at New Green Field School, Saket, New Delhi in 2016, and is currently pursuing his BCA degree from Vivekananda Institute of Professional Studies (VIPS). In 2016, he worked as a Full Stack Developer intern for Xeler8, Inc, in 2017 as Front-end Developer intern for AudienceSutra and as a Full Stack Developer intern for Dhupar Info Tech Pvt Ltd. Currently, he works as a Front-end developer intern for AudienceSutra. He is President for the academic year 2018-19 of ACE – The CSI Student Branch of VIPS.