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 strategies [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 determining the 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 strive for knowing how best to utilize it. 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 represents 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 homogeneous sub-groups within a heterogeneous aggregate market in direct marketing in 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 DSP’s 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.](image1)

![Figure 2. Web user interface process.](image2)
under or over estimation can cause serious business loss or profit [17]. In a changing environment, segmentation helps in determining the 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 bicycle 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 homogeneous sub-groups within a heterogeneous aggregate market in direct marketing in 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 DSPV 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 DSP 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.
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.

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

$N = \text{Number of items}$
$n = \text{Number of user sessions}$

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 centroid 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.

Figure 5. Screenshot of dataset.

Figure 6. Snapshot of 68 boolean predictor variables.

Figure 7. Principal component vs cumulative proportion of variance.

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 the 20 components were taken into consideration.

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]
Market Segmentation Analysis and Visualization using K-Mode Clustering Algorithm...

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, 
\[
M[i,j] = 0, \text{ otherwise}
\]
\( N \) = 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.

The new set of features is sent as an input to the k-mode clustering algorithm. 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 the session id.

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

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].
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 \( x_i = [x_{i1}, x_{i2}, \ldots, x_{in}] \) is described by \( n \) categorical attributes \( A_1, A_2, \ldots, A_n \). Each attribute \( A_j \) describes a domain of values, denoted by

\[
\text{Dom}(A_j) = \{a_{j1}, a_{j2}, \ldots, a_{jn}\}
\]

where \( p_j \) is the number of category values of attribute \( A_j \). The k-modes algorithm operate 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]

\[
P(U, Z) = \sum_{i=1}^{k} \sum_{j=1}^{n} u_{ij} d(x_{ij}, z_{ij})
\]  

Subject to

\[
\sum_{j=1}^{n} u_{ij} = 1, \quad 1 \leq i \leq n
\]

\[
u_{ij} \in \{0,1\}, \quad 1 \leq i \leq n, \quad 1 \leq l \leq k
\]

where

- \( U \) is a \( n \times k \) partition matrix, \( u_{ij} = 1 \), is a binary variable, and \( u_{ij} = 1 \), indicates that object \( x_i \) is allocated to cluster \( C_j \);
- \( Z = \{z_{11}, z_{12}, \ldots, z_{kn}\} \) is a set of vectors representing the centers of the \( k \) clusters,

\[
Z = \{Z_1, Z_2, \ldots, Z_k\} \quad (1 \leq l \leq k)
\]

6.1.2. Distance Function

\[
d(x_{ij}, z_{ij}) \text{ is a distance or dissimilarity measure between object } x_i \text{ and the center of cluster } C_j \text{ on attribute } A_i. \text{ In k-modes algorithm, a simple matching distance measure is used. That is, the distance between two distinct categorical values is 1, while the distance between two identical categorical values is 0. More precisely [47],}
\]

\[
d(x_{ij}, z_{ij}) = f(x) = \begin{cases} 0, & x_{ij} = z_{ij} \\ 1, & x_{ij} \neq z_{ij} \end{cases}
\]

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

1. Problem \( P_1: F_1: Z = \hat{Z}, \text{ solve the reduced problem } P(U, \hat{Z}) \)
2. Problem \( P_2: F_2: U = \hat{U}, \text{ solve the reduced problem } P(\hat{U}, \hat{Z}) \)

Problems \( P_1 \) and \( P_2 \) are solved according to the following two theorems, respectively.

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

\[
u_{ij} = f(x) = \begin{cases} 1, & \sum_{j=1}^{n} d(x_{ij}, z_{ij}) \leq \sum_{j=1}^{n} d(x_{ij}, z_{ij}) \forall jh, 1 \leq h \leq k \\ 0, & \text{otherwise} \end{cases}
\]

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

\[
z_{ij} = a_{ij} \]

where \( a_{ij} \) is a mode of attribute values of \( A_i \) in the cluster \( C_j \) that satisfies

\[
f(a^{(r)}_{ij} | C_j) \geq f(a^{(r)}_{ij} | C_j) \forall r, 1 \leq r \leq p_j
\]

Here \( f(a^{(r)}_{ij} | C_j) \) denotes the frequency count of attribute value \( a_{ij} \) in the cluster \( C_j \), i.e.

\[
f(a^{(r)}_{ij} | C_j) = \frac{u_{ij} | x_{ij} = a_{ij}, u_{ij} = 1 |}
\]

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 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.

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.
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 \( X_i = [x_{i1}, x_{i2}, \ldots, x_{im}] \) is described by \( m \) categorical attributes \( A_1, \ldots, A_m \). Each attribute \( A_j \) describes a domain of values, denoted by

\[
\text{Dom}(A_j) = \{a_{j1}, a_{j2}, \ldots, a_{jp_j}\}
\]

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]

\[
P(U, Z) = \sum_{i=1}^{k} \sum_{x_{ij} \in Z_i} u_{ij} d(x_{ij}, z_{ij})
\]

where

\[
Z = \{Z_1, Z_2, \ldots, Z_k\}
\]

subject to

\[
\sum_{j=1}^{p_j} u_{ij} = 1, \quad 1 \leq i \leq n
\]

\[
u_{ij} \in \{0, 1\}, \quad 1 \leq i \leq n, \quad 1 \leq l \leq k
\]

where

- \( U \) is a \( n \times k \) partition matrix, \( u_{ij} = 1 \) is a binary variable, and \( u_{ij} = 0 \) indicates that object \( x_i \) is allocated to cluster \( C_l \);
- \( Z = \{Z_1, Z_2, \ldots, Z_k\} \) is a set of vectors representing the centers of the \( k \) clusters, where

\[
Z_i = \{Z_{i1}, Z_{i2}, \ldots, Z_{im}\}, \quad (1 \leq l \leq k)
\]

6.1.2. Distance Function

\( d(x_{ij}, z_{ij}) \) is a distance or dissimilarity measure between object \( X_i \) and the center of cluster \( C_l \) 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],

\[
d(x_{ij}, z_{ij}) = f(x) = \begin{cases} 0, & (x_{ij} = z_{ij}) \\ 1, & (x_{ij} \neq z_{ij}) \end{cases}
\]

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

1. Problem \( P_1: F_1(U, Z) = \min P(U, Z) \) Subject to

2. Problem \( P_2: F_2(U) = \min P(U, Z) \) Subject to

Problems \( P_1 \) and \( P_2 \) are solved according to the following two following theorems, respectively.

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

\[
u_{ij} = f(x) = \begin{cases} 1, & \sum_{j=1}^{p_j} d(x_{ij}, z_{ij}) \leq \sum_{j=1}^{p_j} d(x_{ij}, z_{ij}) \forall h, 1 \leq h \leq k \\ 0, & \text{otherwise} \end{cases}
\]

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

\[
z_{ij} = a_{jl},
\]

where \( a_{jl} \) is the mode of attribute values of \( A_j \) in the cluster \( C_l \) that satisfies

\[
f(a_{jl}) \leq f(a_{j}) \forall a \in \text{Dom}(A_j)
\]

where \( f(a_l) \) denotes the frequency count of attribute value \( a_l \) in the cluster \( C_l \), i.e.

\[
f(a_{jl}) \geq nu_{ij} |x_{ij} = a_{jl}, u_{ij} = 1|
\]

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 \( Z(0) \) such that \( P(U(0), Z(0)) \) is minimized. If \( P(U(0), Z(0)) = P(U(0), Z(0)) \), then stop; otherwise go to step 3
3. Determine \( U(0) \) such that \( P(U(0), Z(0)) \) 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 incrementally incremented from 2 to 60 and the model was trained using the k-mode algorithm.

**Algorithm 4:** Evaluation Procedure.

For \( j \) in range (1, 60)

1. Set number_of_clusters = \( j \)
2. Train the model using the k-mode algorithm
3. Obtain cluster centroids
4. 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.

![Figure 8. Number of clusters vs silhouette score.](image)

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 profitability. 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.](image)
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 sub-system 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 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


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 sub-system 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 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 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


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


M. Hartely, "Using Data Mining to Predict Inventory Levels", *Data Mining Book*, 2005.


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 Krekit as a Backend Developer intern. In 2018, he joined Novus 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.