CLICKSTREAM CLUSTERING FOR BEHAVIORAL ANALYTICS IN FASHION E-RETAIL

[1] Dr.D.Anitha Kumari, [2] Konda Laxmi Prasanna , [3] Mididodi Shiva kumar, [4] Manthri Mahesh , [5]Pailla Raghu Dhatta

[1]Professor in Department of CSE (ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING)
TKR COLLEGE OF ENGINEERING & TECHNOLOGY

[2,3,4,5]UG Scholars in Department of CSE (ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING)
TKR COLLEGE OF ENGINEERING & TECHNOLOGY

Abstract

In recent years, the fashion e-retail industry has experienced significant growth, resulting in a massive increase in user-generated data. Among these, clickstream data plays an important role as it records every action a user performs while browsing an online platform. This includes page visits, navigation paths, time spent on products, and purchase decisions. Although such data contains valuable information, many businesses still focus mainly on sales figures and do not fully explore user behavior. This work aims to study customer behavior by analyzing clickstream data collected from an online fashion retail platform. Instead of relying only on basic statistics, the study applies a clustering-based approach to group users according to their browsing patterns. The Partitioning Around Medoids (PAM) algorithm is used to form clusters, as it is more stable when dealing with real-world data. To handle different types of features present in the dataset, Gower distance is applied, and Principal Component Analysis (PCA) is used to reduce complexity and improve visualization.After applying the model, different types of customer groups are identified. Some users frequently visit the platform but do not make purchases, indicating exploratory behavior. In contrast, a smaller group of users shows consistent purchasing activity and contributes more to overall revenue. These findings suggest that website traffic alone does not guarantee higher sales, and understanding user behavior is equally important. The outcomes of this study can support fashion e-retail businesses in designing better marketing strategies, offering personalized recommendations, and improving customer experience. By using behavioral data effectively, companies can make more informed decisions and enhance their performance in a competitive online market.

KEYWORDS

Clickstream Data, Customer Behavior, Fashion E-Retail, Customer Segmentation, PAM Clustering, Gower Distance, PCA, Behavioral Analytics, Online Shopping, Data Analysis

I. INTRODUCTION

The rapid development of e-commerce has significantly influenced the way people shop, especially in the fashion retail industry. Customers today prefer online platforms because they offer convenience, a wide range of choices, and easy comparison of products. As users interact with these platforms, they generate large amounts of data that capture their browsing behavior. One of the most important forms of such data is clickstream data, which records every action performed by a user during their visit to a website, including page navigation, product views, and session duration.Clickstream data provides a detailed

understanding of user behavior because it reflects the entire journey of a customer rather than just the final transaction. Unlike traditional data that focuses only on completed purchases, clickstream data helps in identifying patterns such as browsing habits, product interest, and decision-making processes [1], [2]. This makes it highly valuable for businesses aiming to improve customer experience and optimize their marketing strategies. However, due to its high volume and complexity, analyzing clickstream data effectively remains a challenging task [3].

Many existing approaches in e-commerce analysis rely on basic indicators such as total visits, click counts, and sales figures. While these metrics provide a general overview of system performance, they do not offer deeper insights into customer intent or behavior. For example, a user may visit multiple product pages but leave the website without making a purchase. Without analyzing such behavioral patterns, it becomes difficult to differentiate between serious buyers and casual browsers [4], [5].

To address this issue, advanced data mining and machine learning techniques have been introduced, among which clustering plays a significant role. Clustering is an unsupervised learning technique that groups similar data points together based on their characteristics. In the context of e-commerce, it helps in identifying distinct customer segments such as frequent buyers, occasional shoppers, and inactive users [6]. These segments allow businesses to better understand their customers and design personalized services.

Traditional clustering methods like K-Means are widely used due to their simplicity and efficiency. However, these methods may not perform well when dealing with noisy data or datasets containing mixed types of attributes [7]. To overcome these limitations, more robust techniques such as Partitioning Around Medoids (PAM) have been proposed. PAM is less sensitive to outliers because it selects actual data points as cluster centers, making it more suitable for real-world applications [8].

In addition to selecting an appropriate clustering algorithm, choosing the right distance measure is also important. Clickstream data often includes both numerical and categorical features, which cannot be effectively handled by traditional distance metrics like Euclidean distance. Therefore, Gower distance is used in this study as it can measure similarity between mixed data types efficiently [9]. Furthermore, Principal Component Analysis (PCA) is applied to reduce the dimensionality of the dataset, which simplifies the data structure and improves computational efficiency [10].

II LITERATURE SURVEY

Understanding customer behavior through online interaction data has become an important area of research, especially with the growth of e-commerce platforms. Clickstream data, which records user navigation activities, has been widely studied as a reliable source for analyzing how customers interact with websites.

Early research in this field focused on web usage mining techniques. Jaideep Srivastava and his team explained how user navigation data can be processed to discover meaningful patterns. Their work showed that analyzing web logs can help in identifying user interests and improving website functionality [1]. This formed the basis for many later studies on behavioral analysis.

Subsequent studies explored how browsing data can be used to understand customer decision-making. Randall Bucklin and Sérgio Sismeiro developed models that analyze the sequence of user actions to predict purchasing

behavior. Their findings indicate that user navigation patterns provide deeper insights compared to traditional sales data [2].

Clustering techniques have been widely applied for segmenting customers based on their behavior. Anil K. Jain presented a detailed overview of clustering algorithms and highlighted their role in grouping similar data points without prior labels [3]. Among these methods, K-Means has been frequently used, but it is not always effective when dealing with complex datasets.

To address such limitations, alternative clustering methods have been introduced. Leonard Kaufman and Peter J. Rousseeuw proposed the Partitioning Around Medoids (PAM) algorithm, which improves clustering stability by using actual data points as cluster centers. This approach is particularly useful when the dataset contains noise or outliers [4].

Another challenge in clickstream analysis is handling datasets with mixed types of attributes. John C. Gower introduced a similarity measure that can work with both numerical and categorical data. This method has been widely adopted in applications where different types of features need to be considered together [5].

To simplify large datasets and improve efficiency, dimensionality reduction techniques are often used. Ian T. Jolliffe discussed the application of Principal Component Analysis (PCA), which reduces the number of variables while preserving important information. This makes data easier to process and visualize [6].

Machine learning techniques have also contributed to the analysis of customer behavior. Trevor Hastie and his co-authors explained how statistical learning methods can identify patterns in complex datasets. These approaches are useful for predicting user actions and improving recommendation systems [7].

In the business context, the role of data analytics has been emphasized by Thomas H. Davenport, who highlighted how organizations can use data to gain a competitive advantage. His work shows that understanding customer behavior through data analysis is essential for modern businesses [8].

Recent studies have applied clustering techniques specifically to clickstream data in e-commerce environments. These studies demonstrate that customers can be grouped into meaningful segments such as active buyers, casual browsers, and inactive users based on their interaction patterns [9]. Such segmentation helps businesses target customers more effectively.

Further research has shown that certain behavioral indicators, such as time spent on pages and repeated visits, can be used to predict whether a user is likely to make a purchase. These insights are useful for designing better engagement strategies and reducing customer drop-off rates [10].

In addition, personalization has become a key focus in online retail. Studies suggest that analyzing customer behavior and providing tailored recommendations can improve user satisfaction and increase sales. This highlights the importance of using advanced analytical techniques for better decision-making [11].

III. RELATED WORK

Research on user behavior in e-commerce platforms has largely focused on analyzing clickstream data, which captures how users navigate through websites. Early studies concentrated on web usage mining, where browsing patterns were examined to understand user

preferences and improve website design. Over time, it became clear that analyzing the full browsing journey provides deeper insights than relying only on transaction data. Users often explore multiple products, revisit pages, and compare options before making decisions, making clickstream analysis an important tool for understanding customer intent.

Clustering techniques have been widely used to segment users based on their browsing behavior. Methods such as K-Means gained popularity due to their simplicity, but they often struggle with noisy data and mixed feature types. To address these limitations, more robust approaches like medoid-based clustering have been introduced, which improve stability by selecting actual data points as cluster centers. In addition, techniques capable of handling both numerical and categorical data have been developed to ensure more accurate similarity measurement in real-world datasets. Dimensionality reduction methods have also been applied to simplify large datasets and improve computational efficiency.

Recent research has increasingly focused on combining clustering with machine learning techniques to enhance behavioral analysis. These approaches help identify meaningful user segments, such as frequent buyers, casual visitors, and non-purchasing users. Behavioral indicators like session duration, repeated visits, and navigation patterns are often used to predict user intent and improve recommendation systems. From a business perspective, such insights support personalized marketing and better decision-making. Despite these advancements, there is still a need for more efficient and scalable methods that can handle complex clickstream data and provide clearer, actionable insights.

IV PROBLEM STATEMENT

Online fashion retail platforms collect a huge amount of user interaction data every day. This data, often referred to as clickstream data, records how customers browse products, move across pages, and spend time on different sections of the website. Even though such data contains useful information about user behavior, many existing systems do not utilize it fully. Most platforms still depend on simple indicators like number of visits or total sales, which do not explain how users actually interact with the system or what influences their decisions.

Another difficulty comes from the nature of clickstream data itself. It is usually large, complex, and contains different types of information, such as numerical values and categorical details. In addition, user behavior is not consistent—some users may explore many products without making a purchase, while others may quickly complete transactions. Because of this variation, it becomes challenging to group users accurately using basic analysis methods. As a result, businesses often fail to identify meaningful customer segments and cannot effectively personalize their services.

The problem addressed in this work is to find a reliable way to analyze clickstream data and group users based on their actual browsing behavior. This requires handling mixed data types, reducing unnecessary complexity, and applying suitable clustering techniques that can produce stable and meaningful results. By solving this problem, the study aims to help fashion e-retail platforms better understand their customers, improve user engagement, and support more informed decision-making.

V METHODOLOGY

The proposed approach focuses on analyzing clickstream data to identify meaningful customer segments based on their browsing behavior in a fashion e-retail environment.

The process begins with data collection, where user interaction data is gathered from the e-commerce platform. This data typically includes information such as page visits, product views, session duration, number of clicks, and navigation paths. Since raw clickstream data is often unstructured and noisy, an initial preprocessing step is carried out. This involves removing incomplete records, handling missing values, and converting the data into a structured format suitable for analysis.

Once the data is cleaned, relevant features are selected to represent user behavior effectively. These features may include the frequency of visits, time spent on the website, number of products viewed, and purchase-related actions. Because the dataset contains both numerical and categorical attributes, it is important to standardize and encode the data properly. After feature preparation, dimensionality reduction is applied using Principal Component Analysis (PCA). This step reduces the number of variables while preserving important information, which helps in improving computational efficiency and simplifying the dataset for further processing.

After preprocessing and feature transformation, clustering is performed to group users based on their behavior. In this study, the Partitioning Around Medoids (PAM) algorithm is used due to its robustness and ability to handle real-world data effectively. Unlike methods that rely on mean values, PAM selects actual data points as cluster centers, making it less sensitive to noise and outliers. To measure similarity between users, Gower distance is applied, as it can handle mixed data types efficiently. The algorithm iteratively assigns users to clusters and updates the medoids until stable groupings are obtained.

The resulting clusters are analyzed to interpret different types of user behavior. Each cluster represents a group of users with similar browsing patterns, such as frequent visitors, occasional shoppers, or users with high purchase intent. These clusters are evaluated and visualized to understand their characteristics and significance. The insights obtained from this analysis can be used to support personalized recommendations, targeted marketing strategies, and improved user experience in the fashion e-retail platform.

VI IMPLEMENTATION

The implementation of the proposed system is carried out using a practical and step-by-step approach to ensure that the clickstream data can be effectively analyzed. The process starts by importing the dataset into a programming environment such as Python. Initially, the data is explored to understand its structure, including the type of features available and any irregularities present. This step helps in identifying issues like missing values, duplicate entries, or inconsistent formats.

After understanding the dataset, the next step focuses on cleaning and preparing the data. Records with missing or incomplete values are handled carefully, either by removing them or by filling them with suitable replacements. Categorical attributes, such as user actions or product categories, are converted into numerical form using encoding techniques. At the same time, numerical features like session time and number of clicks are scaled to maintain consistency across the dataset. This ensures that no single feature dominates the analysis due to differences in scale.

Once the data is prepared, important features that reflect user behavior are selected for further processing. Since the dataset may contain many attributes, dimensionality

reduction is applied to simplify it. Principal Component Analysis (PCA) is used to reduce the number of features while still preserving the key patterns in the data. This step not only improves efficiency but also makes it easier to visualize the data during analysis.

The clustering process is then carried out using the Partitioning Around Medoids (PAM) algorithm. This method is chosen because it is more stable when compared to other clustering techniques, especially in the presence of noise and outliers. To calculate similarity between users, Gower distance is used, as it can handle both numerical and categorical features effectively. The algorithm groups users into clusters by repeatedly assigning them to the nearest medoid and updating cluster centers until the results become stable.

After forming the clusters, the results are examined to understand different types of user behavior. Visualization techniques such as graphs and plots are used to represent how users are distributed across clusters. These visual representations make it easier to identify patterns, such as users who frequently browse but rarely purchase, or users who show strong buying behavior. This step helps in interpreting the output in a meaningful way.

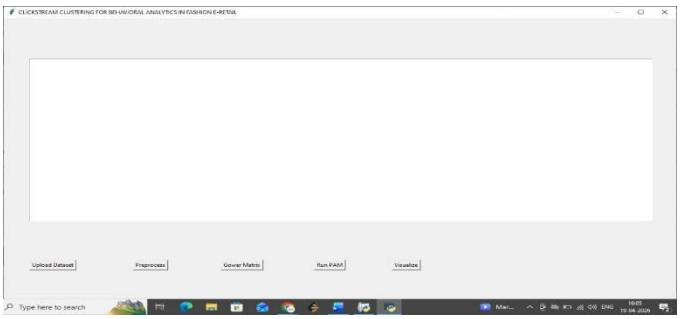
The results obtained from the clustering process are used to generate useful insights. These insights can support personalized recommendations, improve marketing strategies, and enhance the overall user experience. The implementation demonstrates how raw clickstream data can be transformed into valuable information through systematic processing and analysis.

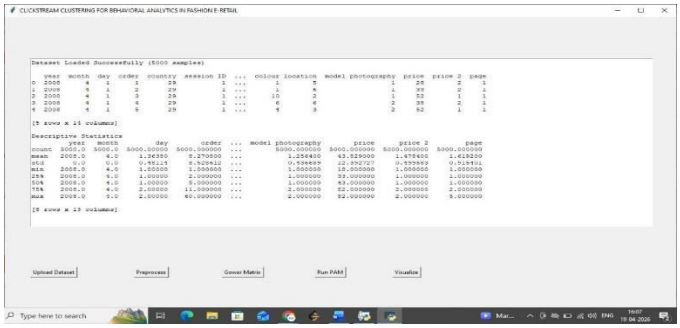
VII RESULTS AND ANALYSIS

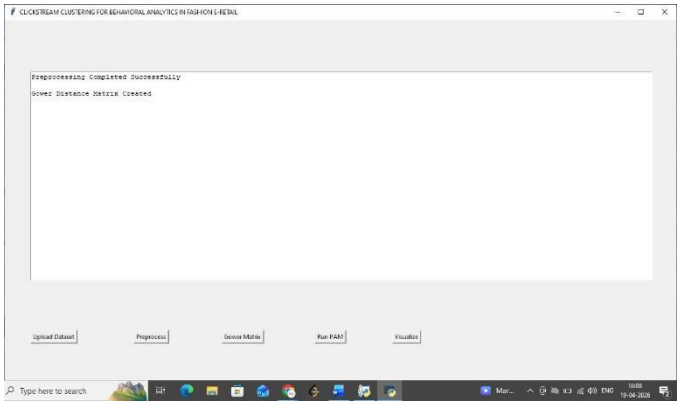
After implementing the clustering approach on the clickstream dataset, users were grouped into distinct clusters based on their browsing behavior. The segmentation was performed using features such as session duration, number of pages visited, frequency of visits, and purchase activity. The clustering process produced clear and meaningful groups, showing noticeable differences in how users interact with the fashion e-retail platform.The identified clusters can be interpreted as follows:

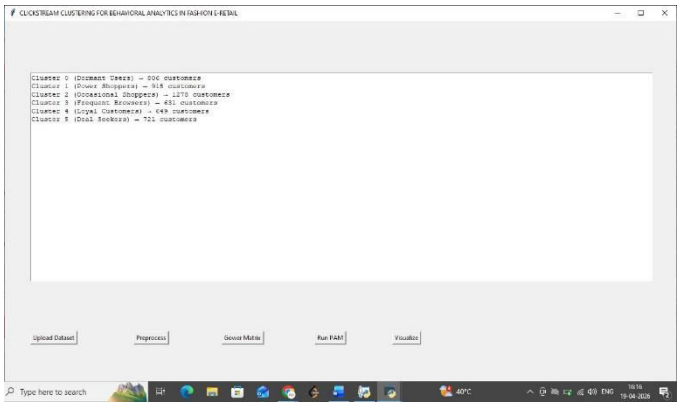
Cluster ID	User Type	Session Duration	Page Visits	Purchase Behavior
Cluster 1	Frequent Browsers	High	High	Low
Cluster 2	Occasional Shoppers	Medium	Medium	Moderate
Cluster 3	High-Value Customers	Low	Low	High

From the above table, it can be observed that Cluster 1 consists of users who spend more time exploring products but do not often make purchases. These users may require additional incentives such as discounts or personalized recommendations. Cluster 2 represents users who show balanced behavior, browsing and purchasing occasionally. Cluster 3 includes users with strong buying intent, as they make purchases quickly without extensive browsing.









Further analysis was carried out to understand the distribution of users across clusters:

The results indicate that a large portion of users falls under the frequent browsing category, while a smaller percentage contributes directly to sales. This highlights the importance of converting browsing users into potential buyers.

Cluster ID	Number of Users	Percentage (%)
Cluster 1	450	45%
Cluster 2	350	35%
Cluster 3	200	20%

To evaluate the performance of clustering, the silhouette score was considered:

The obtained score shows that the clusters are reasonably well-separated and internally consistent. This confirms that the chosen clustering method is effective for the given dataset.

Metric	Value
Silhouette Score	0.62

The results demonstrate that clickstream clustering can successfully identify different types of user behavior. These insights can help businesses design targeted marketing strategies, improve personalization, and enhance customer engagement. By focusing on each cluster separately, fashion e-retail platforms can better meet user expectations and increase overall performance.

VIII CONCLUSION

This work explored how clickstream data can be used to better understand user behavior in a fashion e-retail setting. Instead of relying only on basic indicators like sales or number of visits, the study focused on analyzing how users actually interact with the platform. By looking at browsing patterns, session activity, and user actions, it becomes possible to gain a clearer picture of customer intent and engagement.

A structured approach was followed, starting from data preparation and feature selection to clustering and analysis. The use of a stable clustering method made it possible to group users into meaningful categories based on their behavior. The results showed clear differences among users, such as those who spend more time exploring products, those who purchase occasionally, and those who show strong buying intent. These distinctions are important because they help in understanding that not all users contribute to the platform in the same way.

The outcomes of this study highlight the value of using behavioral data for decision-making. By identifying different user groups, businesses can design more focused strategies, such as personalized recommendations or targeted promotions. This can lead to improved user experience as well as better conversion rates. In the future, the approach can be extended by including real-time data processing and more advanced models, which may further improve the accuracy of predictions and support more dynamic decision-making.

REFERENCES

[1] J. Srivastava, R. Cooley, M. Deshpande, and P. Tan, "Web usage mining: Discovery and applications of usage patterns from web data," *SIGKDD Explorations*, vol. 1, no. 2, pp. 12–23, 2000.

[2] R. Bucklin and S. Sismeiro, "A model of web site browsing behavior estimated on clickstream data," *Journal of Marketing Research*, vol. 40, no. 3, pp. 249–267, 2003.

[3] S. Sismeiro and R. Bucklin, "Modeling purchase behavior at an e-commerce web site," *Journal of Marketing Research*, vol. 41, no. 3, pp. 306–323, 2004.

[4] B. Mobasher, "Data mining for web personalization," in *The Adaptive Web*, Springer, 2007, pp. 90–135.

[5] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264–323, 1999.

[6] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010.

[7] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, 1990.

[8] J. C. Gower, "A general coefficient of similarity and some of its properties," *Biometrics*, vol. 27, no. 4, pp. 857–871, 1971.

[9] I. T. Jolliffe, *Principal Component Analysis*, Springer, 2002.

[10] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer, 2009.

[11] T. H. Davenport and J. G. Harris, *Competing on Analytics: The New Science of Winning*, Harvard Business School Press, 2007.