THE 2nd INTERNATIONAL CONFERENCE ON
"GREEN SOLUTIONS AND EMERGING TECHNOLOGIES
FOR SUSTAINABILITY" (GSETS 2025)
Email: gsets@hutech.edu.vn
Website: https://www.hutech.edu.vn/gsets2025

Enhancing E-Commerce Store Clustering Using Frequent Itemset Mining and Mixed Data Analysis

Dat Nguyen[1], Huong Bui[1*]
[1]Faculty of Information Technology, HUTECH University, Ho Chi Minh City, Vietnam
**Corresponding Author: (Phone: + 84 909 344 837; Email: bd.huong@hutech.edu.vn)*

Abstract

Clustering stores on e-commerce platforms is an important task for performance analysis and business strategy optimization. This paper suggests a new way to do things that uses frequent itemset mining to add binary features to the clustering process. This makes the data representation richer and the clustering more accurate. The paper provides a comprehensive outline of the data preprocessing steps, ensuring high-quality input data before feature integration. We evaluate the performance of three widely-used clustering algorithms—K-Means, Gaussian Mixture Model, and Agglomerative Clustering—both before and after the introduction of frequent itemset-based features. The results show that the added features uncover hidden relationships within the data, leading to improved clustering precision. The paper also looks at related methods, like K-Prototypes and Gower distance, that can help you judge the accuracy of clustering results when the data has both continuous and binary variables. The results show that using frequent itemset mining makes the clustering process much better, creating clusters that are more distinct, cohesive, and well-separated. These improved clusters provide deeper insights for identifying important store segments and suggest tailored business strategies for each group, thus contributing to enhanced business performance. The proposed method enables managers to better identify critical store clusters, offering actionable insights for developing strategies that enhance customer satisfaction and drive revenue growth across the e-commerce platform.

Keywords: *Data Preprocessing, E-commerce Store Clustering, Frequent Itemset Mining, Mixed Data Analysis, Unsupervised Learning.*

1. INTRODUCTION

E-commerce is increasingly becoming an indispensable part of the global economy, especially with the rapid growth of online platforms. In this context, clustering stores data on e-commerce platforms plays a crucial role in optimizing business operations and enhancing platform competitiveness. Many studies have focused on applying data mining techniques to effectively classify and segment customers in this domain.

In e-commerce, the research by Rachid et al. (2018) combined the LRFM model with the K-means algorithm to classify and detect early signs of customer churn. Kamthania et al. (2018) used PCA and K-Modes to cluster customers to better understand the shopping behavior and geographical distribution of users, thereby supporting small businesses and startups to build effective business strategies. In business management, Van Leeuwen & Koole (2022) focused on the application of unsupervised learning to segment customers in the hotel industry to optimize marketing strategies and enhance customer experience. In data analytics, Bhattacharjee & Mitra (2021) synthesized and classified density-based clustering algorithms, emphasizing applications in areas such as geography and multimedia data. Patel & Kushwaha (2020) compared K-Means and Gaussian Mixture Models (GMM) in cloud workload specifications, showing that GMM models resource details better, while K-Means is faster but suitable for simple data.

While clustering studies focused on grouping and classifying data, frequent itemset mining methods complemented and extended these analyses by extracting deep relationships. For example, Bikku (2018) proposed a weight-based itemset mining method to improve the accuracy in real-time e-commerce data. This method helped to select typical products for creating combined samples, which is effective in analyzing customer behavior and optimizing business strategy. A study by Jamshed et al. (2020) used a deep learning model that combined CNN and LSTM to find sequential patterns in a dynamic database. They found that adding Discrete Wave Transform (DWT) improved accuracy, performance, and scalability while reducing processing time. Bashir (2020) suggested the FT-PatternGrowth algorithm as a better way to mine frequent fault-tolerant item sets. It uses FP-tree and a divide-and-conquer strategy to speed up processing and cut costs, achieving high performance in big data analysis with a low support threshold.

Research on frequent set mining has demonstrated the ability to extract important patterns from data and has opened up the potential to use these features to improve the accuracy and meaningfulness of clustering algorithms. Kumar et al.(2022) combined association rule mining and machine learning to analyze websites, optimize conversion rates, reduce cart abandonment, improve user experience, and support business strategies like product personalization and search enhancement. Sinthuja et al.(2024) introduced the CL-LP-MAX-tree method, which combines clustering and a linear prefix tree to mine maximal frequent item sets more efficiently, reducing running time, saving memory, and being suitable for big data in e-commerce. A study by Rouane et al.(2019) combined clustering and frequent itemset mining to improve biomedical text summarization, using K-Means and Apriori to select key content. This method outperformed other summarization techniques.

Although there have been many studies combining clustering and frequent itemset mining in different fields, there are still significant limitations. Most of the studies focus on specific types of data without really exploring the specific context of e-commerce stores, especially in Southeast Asia and Vietnam, such as Tiki. Also, the current methods still rely on old methods like Apriori and FP-Tree, not fully utilizing modern tools or systematically combining frequent itemset mining and clustering. This limits the accuracy and applicability of clustering results in practice.

This study aims to apply FP-MAX to improve the clustering efficiency of stores on the Tiki platform. The main goal is to create the best way to combine FP-MAX with three well-known clustering algorithms (K-Means, agglomerative clustering, and Gaussian mixture model) so that analysis is more accurate and faster. In addition, the study also focuses on in-depth analysis of practical solutions, supporting stores on Tiki to plan effective business strategies. The results not only contribute scientific significance but also bring high application value, helping Tiki and similar platforms increase their competitiveness in the vibrant e-commerce market.

The rest of the paper is organized as follows. Section 2 presents the proposed method, including the problem statement, data collection and preprocessing, FP-Max algorithm, evaluation method, and implementation model. Section 3 presents the clustering and cluster analysis results. Section 4 concludes the paper.

2. PROPOSED METHOD

2.1 OBJECTIVE

In this paper, we present a new way for e-commerce stores to use clustering and maximal frequent itemset mining together to add more data features and make clustering better. Specifically, we employed three algorithms: K-Means (Yuan & Yang, 2019), Agglomerative Clustering (Tokuda et al., 2022) and Gaussian Mixture Models (Ban et al., 2018) for clustering due to their versatility and effectiveness. Additionally, we utilize the FP-Max (Ziani & Ouinten, 2009) to extract maximal frequent itemsets, thereby enriching the dataset with binary features derived from meaningful patterns. The main goal of this study is to show that clustering algorithms work better when they are combined with maximal frequent itemset mining techniques. The process comprises the following six steps: data collection, data preprocessing, FP-Max mining, clustering before and after FP-Max combination, model evaluation, and cluster analysis. We will discuss each step in detail below.

2.2 DATA COLLECTION

The data used in this study was collected from the Tiki e-commerce platform, one of the leading e-commerce platforms in Vietnam. A total of 6 features and 933 stores on this platform were collected and analyzed. Description of the dataset in Table 1 .

Table 1: Dataset description

Features	**Description**
Revenue	Estimated revenue of each store in 1 month
YearJoined	Number of years selling on the Tiki platform
Followers	Number of store followers
ChatResponse	The store's chat response rate
RatingQuality	Average store rating quality
PositiveQuality	Quality of the store's average positive review rate

2.3 DATA PREPROCESSING

Data preprocessing is a crucial step to ensure that the data used for analysis or model building is of high quality and can yield accurate results. In this study, we performed the following preprocessing steps:

Data Cleaning: To ensure accuracy and consistency and retain unique entries, invalid data rows, including those with inconsistent or missing values and duplicate records, were identified and removed.

Data Normalization: We applied the StandardScaler method to scale features uniformly. This enhances machine learning model efficiency and improves analysis accuracy.

Noise Detection and Handling: Initially, we used a boxplot to check for noise and applied Winsorization to reduce outlier impact while preserving data integrity. Then we used the Isolation Forest to isolate noisy data and remove them. Finally, the DBSCAN (Deng, 2020) was applied to identify and eliminate remaining noise from the previous steps by grouping high-density data points and removing points outside these groups. Handling outliers helps increase the accuracy of the machine learning model.

2.4 FP-MAX

In this study, the initial dataset consisted of continuous variables describing the business characteristics of stores on the Tiki e-commerce platform. Two important data transformation steps were performed to enable the application of the FP-MAX algorithm and ensure the data met its requirements. First, the continuous variables were discretized into *Low*, *Medium*, and *High* value ranges. Next, discretized data were converted into binary form, in which each variable was represented as three binary attributes corresponding to each value range.

Specifically, a continuous variable $\boldsymbol{X}$ is transformed into three binary attributes $\boldsymbol{X_{\{Low\}}}$, $\boldsymbol{X_{\{Medium\}}}$, $\boldsymbol{X_{\{High\}}}$ where the value 1 indicates that the value of $\boldsymbol{X}$ belongs to the corresponding range, and the value 0 otherwise. This approach helps the input data meet the required format for the FP-Max. Subsequently, the FP-Max was applied to mine the maximal frequent itemsets for tense consistency and improved flow.

The process of mining the maximal frequent itemset includes three main steps. First, the frequent itemset $\boldsymbol{I = \{i_1, i_2, \dots, i_k\}}$ is identified based on its support of $\mathbf{Support(\boldsymbol{I})}$ in the dataset $\boldsymbol{D}$. The condition for $\boldsymbol{I}$ to be considered frequent is:

$$\text{Support}(I) = \frac{\text{Transaction number containing } I}{\text{Total transactions}} \geq \text{min\_support} \tag{1}$$

In which, **min_support** is the user-defined minimum support threshold.

Second, an itemset I is called a maximal frequent itemset if:

$$\forall X \supset I: \text{Support}(X) < \text{min\_support} \tag{2}$$

In other words, an itemset I is maximal when none of its superset is frequent itemset. This property helps eliminate unnecessary sub-itemsets and reduces the number of results that need to be stored. After extracting the maximal itemsets using FP-Max, the itemsets are converted into binary features, 1 if the itemset condition is satisfied, otherwise 0.

2.5 EVALUATION METHOD

To evaluate the performance of the clustering model in our study, we used three popular clustering indices: Silhouette Score (Shutaywi & Kachouie, 2021), Davies-Bouldin Index (Petrovic, 2006), and Calinski-Harabasz Index (Wang & Xu, 2019). These indices allow us to assess the quality of clusters based on the cohesion within clusters and the separation between clusters.

2.6 MODEL IMPLEMENT

We used the Silhouette method with each algorithm and tested cluster numbers ranging from 3 to 10 to determine the optimal number of clusters for the clustering process. The results in Figure 1 show that the model achieved the best clustering efficiency when using K-Means with the number of clusters being 5. Similarly, the optimal number of clusters was determined to be 4 for Agglomerative in Figure 2 and 3 for Gaussian Mixture Model in Figure 3. These cluster numbers were then applied to perform clustering with each corresponding model.

We used the Forward Selection method to sequentially add new binary features created from the FP-Max process into the clustering models K-Prototypes, Agglomerative, and Gaussian Mixture Models. If a new binary feature positively impacted the improvement of evaluation metrics, it was retained; otherwise it was discarded. The Gower distance (Tuerhong & Kim, 2014) was also applied to the K-Prototypes (Ji et al., 2013) and Agglomerative algorithms to enhance the evaluation of the mixed dataset.

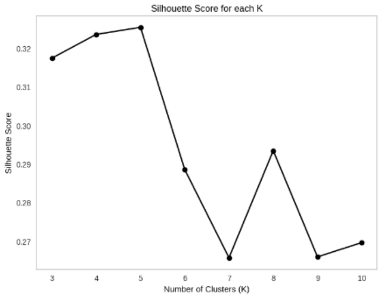

Figure 1: Optimal cluster number for K-Means

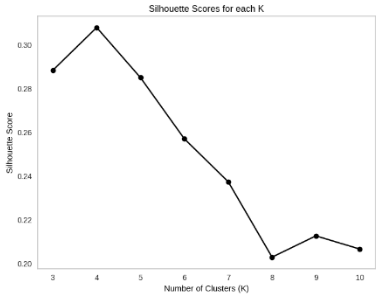

Figure 2:Optimal cluster number for Agglomerative

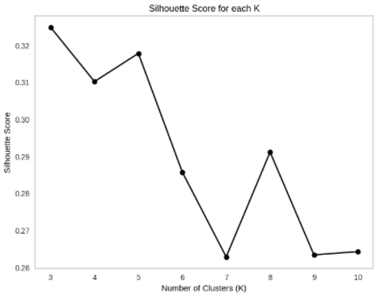

Figure 3: Optimal cluster number for Gaussian Mixture Model

3. EXPERIMENTAL RESULTS

3.1 COMPARE THE RESULTS BEFORE AND AFTER ADDING THE BINARY VARIABLES

The research results are divided into two stages: before and after integrating binary features created from maximal frequent itemset mining. Before integration, the clustering process was based only on continuous variables describing the business characteristics of stores on the Tiki platform. The clustering efficiency was limited in Table 2, and the visual results in Figure 4, 5, 6 showed that the clusters did not achieve a good level of connectivity, reducing the analysis efficiency.

Table 2: Evaluation of clustering result metrics

Algorithms	Silhouette	Davies-Bouldin	Calinski-Harabasz
K-Means	0.325	1.088	415.935
Agglomerative Clustering	0.308	1.036	361.240
Gaussian Mixture Model	0.325	1.249	378.246

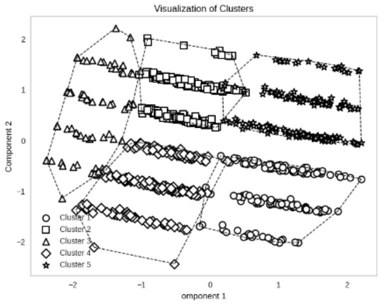

Figure 4: Visualization of K-Means results

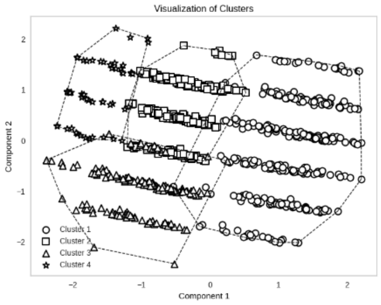

Figure 5: Visualization of Agglomerative results

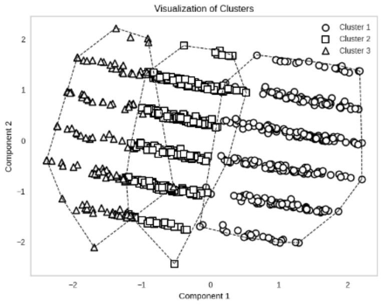

Figure 6: Visualization of Gaussian Mixture Model results

After integrating new binary features created from FP-Max, the clustering efficiency improved significantly, as shown in the evaluation results in Table 3. Algorithms such as K-Prototype, Agglomerative Clustering, and Gaussian Mixture Model achieved notable improvements. Figures 7-9 also show better cohesion and separation between clusters.

Table 3: Evaluation of clustering result metrics after combined new binary features

Algorithms	Silhouette	Davies-Bouldin	Calinski-Harabasz
K-Means	0.459	0.963	495.393
Agglomerative Clustering	0.553	0.628	1451.059
Gaussian Mixture Model	0.341	1.234	373.190

The comparison between the two stages clearly shows the advantages of integrating binary features. The results emphasize that maximal frequent item set mining not only enriches the feature space but also significantly improves the clustering quality, creating a stronger foundation for developing effective business strategies to support e-commerce platforms such as Tiki in strengthening their competitiveness in the market.

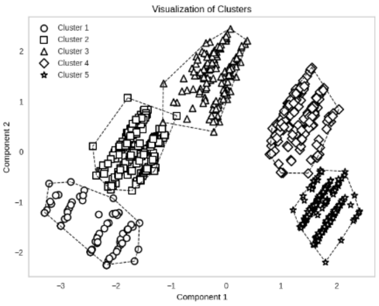

Figure 7: Visualization of K-Means results after combined new binary features

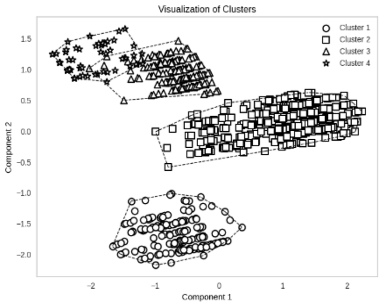

Figure 8: Visualization of Agglomerative results after combined new binary features

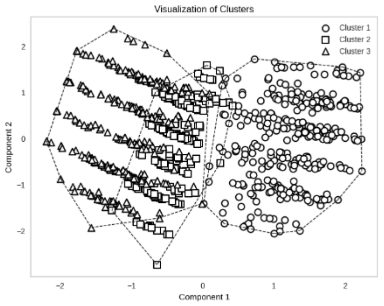

Figure 9: Visualization of Gaussian Mixture Model results after combined new binary features

3.2 CLUSTER ANALYSIS

Since the K-Means algorithm determined the largest number of clusters among the three algorithms, it was chosen for detailed analysis. A larger number of clusters allows K-Means to identify store groups with distinct business characteristics and service quality in greater detail.

Cluster 1: Long-standing stores with low business efficiency. Revenue, followers, and positive feedback are very low, as is service quality. Comprehensive improvements are needed in product quality, service, promotions, and customer engagement.

Cluster 2: Newer stores with good business efficiency but varying metrics, indicating uneven performance. Platforms should focus on optimizing strategies, promoting products, and ensuring consistent service quality.

Cluster 3: Long-standing stores operating inefficiently, with low metrics but some positive scores. Improvements in operations, product innovation, and customer understanding can unlock growth potential.

Cluster 4: Relatively new stores with average performance. Feedback and service quality are stable but not high. Growth can be achieved through training programs, sales support tools, and better customer service.

Cluster 5: Long-standing, highly efficient stores. Revenue varies, but quality scores and feedback are superior. These stores exemplify business success and should be retained and supported with marketing and partnership programs.

4. CONCLUSION

The results of the study have shown the importance of integrating binary features mined from the maximal frequent itemset into the clustering process. Significant improvements in metrics such as Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index after integrating these features show that they help to discover hidden relationships in the data that were previously undetectable with continuous variables alone. This emphasizes the complementary role of frequent itemset mining in enriching the feature space, especially for datasets with diverse features. These findings not only provide a solid basis for applying the same method on other e-commerce platforms but also suggest the potential for integrating weighted pattern mining techniques in the future. This method can help to discover patterns that are more important in shaping business or customer behavior, thereby optimizing clustering strategies. This is especially useful in large and complex data contexts where important patterns may be missed using traditional methods alone. Applying this technique can open up new research directions, improve clustering quality, and make more valuable contributions to the e-commerce field.

REFERENCES

Ban, Z., Liu, J., & Cao, L. (2018). Superpixel Segmentation Using Gaussian Mixture Model. *IEEE Transactions on Image Processing*, *27*(8), 4105–4117. https://doi.org/10.1109/TIP.2018.2836306

Bashir, S. (2020). An efficient pattern growth approach for mining fault tolerant frequent itemsets. *Expert Systems with Applications*, *143*, 113046. https://doi.org/10.1016/j.eswa.2019.113046

Bhattacharjee, P., & Mitra, P. (2021). A survey of density based clustering algorithms. *Frontiers of Computer Science*, *15*(1), 151308. https://doi.org/10.1007/s11704-019-9059-3

Bikku, T. (2018). A new weighted based frequent and infrequent pattern mining method on realtime E-commerce. *Ingénierie Des Systèmes d'information*, *23*(5), 121–138. https://doi.org/10.3166/isi.23.5.121-138

Deng, D. (2020). DBSCAN Clustering Algorithm Based on Density. *2020 7th International Forum on Electrical Engineering and Automation (IFEEA)*, 949–953. https://doi.org/10.1109/IFEEA51475.2020.00199

Jamshed, A., Mallick, B., & Kumar, P. (2020). Deep learning-based sequential pattern mining for progressive database. *Soft Computing*, *24*(22), 17233–17246. https://doi.org/10.1007/s00500-020-05015-2

Ji, J., Bai, T., Zhou, C., Ma, C., & Wang, Z. (2013). An improved k-prototypes clustering algorithm for mixed numeric and categorical data. *Neurocomputing*, *120*, 590–596. https://doi.org/10.1016/j.neucom.2013.04.011

Kamthania, D., Pawa, A., & Madhavan, S. (2018). Market Segmentation Analysis and Visualization using K-Mode Clustering Algorithm for E-Commerce Business. *Journal of Computing and Information Technology*, *26*(1), 57–68. https://doi.org/10.20532/cit.2018.1003863

Kumar, B., Roy, S., Sinha, A., Iwendi, C., & Strážovská, Ľ. (2022). E-Commerce Website Usability Analysis Using the Association Rule Mining and Machine Learning Algorithm. *Mathematics*, *11*(1), 25. https://doi.org/10.3390/math11010025

Patel, E., & Kushwaha, D. S. (2020). Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model. *Procedia Computer Science*, *171*, 158–167. https://doi.org/10.1016/j.procs.2020.04.017

Petrovic, S. (2006, October). A comparison between the silhouette index and the davies-bouldin index in labelling ids clusters. In Proceedings of the 11th Nordic workshop of secure IT systems (Vol. 2006, pp. 53-64). Citeseer.

Rachid, A. D., Abdellah, A., Belaid, B., & Rachid, L. (2018). *Clustering Prediction Techniques in Defining and Predicting Customers Defection: The Case of E-Commerce Context*. *8*(4).

Rouane, O., Belhadef, H., & Bouakkaz, M. (2019). Combine clustering and frequent itemsets mining to enhance biomedical text summarization. *Expert Systems with Applications*, *135*, 362–373. https://doi.org/10.1016/j.eswa.2019.06.002

Shutaywi, M., & Kachouie, N. N. (2021). Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering. *Entropy*, *23*(6), 759. https://doi.org/10.3390/e23060759

Sinthuja, M., Pravinthraja, S., Dhanalakshmi, B. K., Gururaj, H. L., Ravi, V., & Jyothish Lal, G. (2024). An efficient and resilience linear prefix approach for mining maximal frequent itemset using clustering. *Journal of Safety Science and Resilience*, S2666449624000689. https://doi.org/10.1016/j.jnlssr.2024.08.001

Tokuda, E. K., Comin, C. H., & Costa, L. D. F. (2022). Revisiting agglomerative clustering. *Physica A: Statistical Mechanics and Its Applications*, *585*, 126433. https://doi.org/10.1016/j.physa.2021.126433

Tuerhong, G., & Kim, S. B. (2014). Gower distance-based multivariate control charts for a mixture of continuous and categorical variables. *Expert Systems with Applications*, *41*(4), 1701–1707. https://doi.org/10.1016/j.eswa.2013.08.068

Van Leeuwen, R., & Koole, G. (2022). Data-driven market segmentation in hospitality using unsupervised machine learning. *Machine Learning with Applications*, *10*, 100414. https://doi.org/10.1016/j.mlwa.2022.100414

Wang, X., & Xu, Y. (2019). An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index. *IOP Conference Series: Materials Science and Engineering*, *569*(5), 052024. https://doi.org/10.1088/1757-899X/569/5/052024

Yuan, C., & Yang, H. (2019). Research on K-Value Selection Method of K-Means Clustering Algorithm. *J*, *2*(2), 226–235. https://doi.org/10.3390/j2020016

Ziani, B., & Ouinten, Y. (2009). Mining maximal frequent itemsets: A java implementation of FPMAX algorithm. *2009 International Conference on Innovations in Information Technology (IIT)*, 330–334. https://doi.org/10.1109/IIT.2009.5413790