DOI: 10.15439/2023F1377
ISSN 2300-5963 ACSIS, Vol. 35

Clusterization methods for multi-variant e-commerce interfaces

Adam Wasilewski
0000-0002-1653-5005
Wroclaw University of Science and Technology
Wyb. Wyspianskiego 27, 50-370 Wroclaw, Poland
Fast White Cat S.A., Poland
Email: adam.wasilewski@pwr.edu.pl

***Abstract*—E-commerce is a very popular method that lets consumers to purchase goods and services. The ability to purchase items online has increased the need for effective recommendation systems. Such recommendations usually relete to products in which the customer may be interested. However, there are wider opportunities to tailor e-commerce to individual customer needs and behaviour. In this paper, the architecture of the e-commerce platform (named AIM^2), which allows providing a dedicated interface to selected user groups, is discussed. A key component of the platform is the module responsible for dividing customers into groups, using selected clustering methods. Each of the implemented methods can be parameterised to adapt the customer segmentation to a given e-commerce business owner's requirements. This article describes the results of an analysis of the impact of selected methods and parameters on clustering results. Moreover, it identifies key metrics that should be considered when selecting clustering conditions during the implementation of the platform. Finally, the main results of the pilot implementation of AIM^2 are presented to assess the effectiveness of the multi-variant user interface.**

I. Introduction

ONE major drawback of existing e-commerce systems is that they display little ability to take into account differences in the users' knowledge, style, and preferences. Meanwhile, users of the systems are different, and the interfaces served to them could be different. The concept of AUI (Adaptive User Interfaces) is increasingly frequent in implementation in the modern IT systems, but this trend is less visible in e-commerce. AIM^2 platform, described in the paper, is an example of the practical implementation of the AUI concept in e-commerce systems. Its aim is to serve dedicated e-commerce user interfaces based on user groups defined by clustering using AI methods, monitoring the results and optimizing the solution.

Personalized web-based system user interfaces can be defined as systems that can automatically adjust their presentation, content, and structure based on the user's characteristics, needs, or preferences [12]. Such solutions can improve the usability and effectiveness of the interface by its adaptation to the user's behaviour. It may also reduce the user's cognitive load, which in turn reduces the user's tiredness and the number of committed errors. Better UX may increase the user's satisfaction and motivation. Design and implementation of AUI may be a complex task, requiring usage of interdisciplinary solutions [15]. Such system may be based on multi-agent infrastructure [3] but nowadays, great opportunities are offered by AI methods as well.

Among the various existing approaches to creating self-adaptive user interfaces is the use of AI-based clustering to divide customers into groups and serve these groups with a dedicated interface. AI clusterization methods leverage advanced algorithms to discover hidden patterns, structures, or relationships within datasets. Information about users' activities, events, purchases and other factors that should potentially affect the interface can be used to group e-commerce users. If designated user groups are served with an interface tailored to their behaviour, an increase in the e-commerce performance can be expected, which should ultimately have a positive impact on profitability in business terms.

In Section 2 previous works on using clusterizaton methods in e-commerce are described. Section 3 briefly shows the architecture of AIM^2 platform and research methodology. Experiments related to clusterization methods and their parameters are detailed and discussed in Section 4. Section 5 concludes the work.

II. Related Work

Clusterization methods are an effective way to group similar items together and provide personalized recommendations. However, the choice of method depends on the specific requirements and limitations of the e-commerce interface. Commonly used clusterization methods are:

- Hierarchical methods, which create a hierarchical structure of clusters by iteratively merging or splitting clusters, e.g. agglomerative clustering and divisive clustering;
- Partitioning methods, which divide the dataset into a predetermined number of clusters, e.g. K-means, K-medoids, and Fuzzy C-means clustering;
- Density-based methods, which focus on identifying regions of high-density data points and separating them from sparse regions, e.g. DBSCAN (Density-Based Spatial Clustering of Applications with Noise);

Project "Self-adaptation of the online store interface for the customer requirements and behaviour" co-funded by the National Centre for Research and Development under the Sub-Action 1.1.1 of the Operational Programme Intelligent Development 2014-2020.

- Spectral, which utilizes the concept of spectral graph theory, which relates the eigenvectors and eigenvalues of a similarity matrix to the underlying structure of the data;
- Model-based methods, which assume that the data points are generated from a statistical model or distribution, e.g. Gaussian Mixture Models (GMM) and Expectation-Maximization (EM) algorithm.

Almost all of the methods mentioned were verified for their application for e-commerce solutions.

Hierarchical clustering in e-commerce applications is discussed in [18]. It is worth noting that the standard algorithm for hierarchical agglomerative clustering (HAC) can be slow even for medium data sets [1]. Research related to the record linkage system for e-commerce products was also conducted using this approach [9]. HAC algorithm used for customer segmentation was described in [10] and the findings could also be used to segment e-commerce customers.

K-means algorithm can be used for segmenting e-commerce customers to obtain groups of customers with different characteristics [5]. In the clustering customer purchase data can be taken into account [17]. According to researchers K-means clustering is quite efficient algorithm. However, its computational difficulty is influenced by the size of the dataset, the number of clusters, and the initialization of the cluster centroids. K-medoids (uses the medoid instead of the mean) was used for e-commerce customer segmentation [21]. Classification method of e-commerce user behavior based on Fuzzy C-Means Clustering was proposed to improve the clustering analysis effect of e-commerce user behavior in [20].

DBSCAN [8] was used to process uneven density data based on information from e-commerces [19]. DBSCAN algorithm seems to be more advantageous when compared to other algorithms because doesn't need one to specify the number of clusters within the knowledge a priori [13].

Some researchers have also compared the effectiveness of different clustering algorithms used in e-commerce solutions. A comparison of the performance of the K-means and DBSCAN (Density-based spatial clustering of applications with noise) algorithms in e-commerce applications indicated slightly better results obtained by the second method [2]. In another study, an accuracy comparison indicated an advantage for DBSCAN over K-means [7]. DBSCAN seems to correspond more to human intuitions of clustering, rather than distance from a central clustering point (e.g. K-means) [13].

In some e-commerce applications spectral clustering can be an effective clustering method [6]. There is also the possibility of using spectral clustering in conjunction with K-means clustering [22].

There are relatively few publications on the application of the GMM method [23] to e-commerce-related analysis. Some possibilities for applying this method in product recommendation are indicated [11]. According to [16] the K-means method has lower computational requirements, and could potentially yield clustering results similar to those of the GMM method.

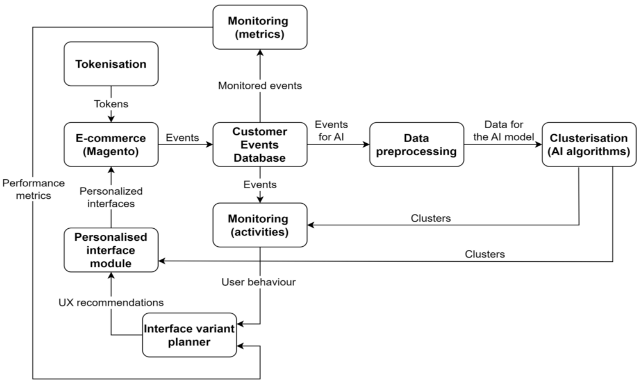

Fig. 1. AIM^2 Business Architecture

III. AIM^2 ARCHITECTURE

The AIM^2 business architecture contains the core modules of the system and key integration interfaces (Fig. 1). It includes:

- Tokenization - responsible for ensuring the anonymity of the data collected by the platform;
- E-commerce - Adobe Magento-based e-shop with implementation of PWA (Progressive Web App) technology;
- Customer Events Database - e-commerce customer behaviour data storage;
- Preprocessing - the module to prepare data to analyse and to decrease the time required to generate clusters;
- Clusterisation - analyses the information collected about the behaviour of e-commerce customers and to divide them into groups characterised by similar shop use;
- Monitoring - identifies user patterns that can be used to design variants of dedicated interfaces and verifies the performance of the interface variants;
- Interface management - supports variants definitions shown to selected users;
- Self-adaptation algorithm - automatically implement micro changes to the interface and accept or reject them depending on the impact on e-commerce performance metrics.

The interface adaptation process starts with initial clustering of e-commerce customers, using learning data from a possible long period. One of the clusterization effects is a set of customer groups that are heterogeneous based on certain characteristics. The AIM^2 platform has four clustering methods implemented: agglomerative clustering, the K-means method, DBSCAN and spectral clustering. Some of them can be applied with different parameter values, e.g.:

- model (type=string, default='kmeans') algorithm used - k-means ('kmeans'), agglomerative clustering ('agg'), DBSCAN ('dbscan') and spectral ('spectral')
- pca (type=float, default=0.999) - the amount of data variance retained is greater than the percentage specified in the parameter

- init (type=string, default='k-means++') - method for initialising cluster centres for the k-means algorithm:
  - 'k-means++' samples the dataset and counts k-means on it [4],
  - 'random' randomises the centres.
- max_iter (type=integer, default=300) - maximum number of iterations in the k-means algorithm

The choice of clustering module parameter settings affects the allocation of customers into clusters and, consequently, the interface variants that will be provided to them. For this reason, it is important to properly tune this module and match the requirements of a particular e-commerce. The research conducted was aimed at verifying the outflow of different values of clustering parameters on its effects. The quality of clustering can be considered in two aspects - the objective one, resulting from classical methods of cluster evaluation (as: Silhouette score, entropy, Calinski-Harabasz score and Davies-Bouldin score) and contextual, resulting from the requirements for clusters intended to be the basis for providing different variants of the e-commerce customer interface. During the research, experiments were carried out on a dataset covering a period of 4 months (548.922 e-commerce user sessions) to verify the impact of the choice of clustering method and the parameters (pca, init, max_iter) on the distribution of customers in the clusters. From a business point of view, clusters should be of similar size, since it makes most sense to prepare a dedicated interface variant. There should not be too many clusters (groups of customers), because the creation of an interface variant is a non-zero cost, so from a business point of view, it is preferable to use methods that generate clusters with a size of no less than X% of the population (X can be treated as a parameter and vary depending on the specific e-commerce, for the purposes of the study X=5 was assumed).

IV. Research results

The research was carried out in two stages. In the first part, clustering was performed with all four methods available on the AIM^2 platform (Tab. I).

The following were analysed: clustering time, number of clusters, size of the largest and smallest cluster. From the point of view of the adopted business requirements, the clustering time (without pre-processing) should be as short as possible, the number of clusters not exceeding 10 and the size of the smallest cluster not less than 5% of the total number of customers (91.819 e-commerce users were clustered in the experiment).

K-means and DBSCAN clustering were found to be significantly more time-efficient than the other methods. Nevertheless, the duration of the longest clustering (using the spectral method) would also be acceptable, as it would be feasible in the time allowed for cyclic updating of the cluster composition in AIM^2.

The number of clusters obtained, in particular as many as 282 clusters generated using the DBSCAN method, is noteworthy. This is well beyond the upper limit of the number of clusters taken as a business requirement. The lack of predictability of the number of clusters and the inability to reduce the number of clusters make the DBSCAN method practically useless from the point of view of the self-adaptation mechanism of the e-commerce interface implemented in AIM^2. The other three methods had the number of clusters set to 10. This number, from a business point of view, seems to be the maximum number of interface variants that should be served to e-commerce customers.

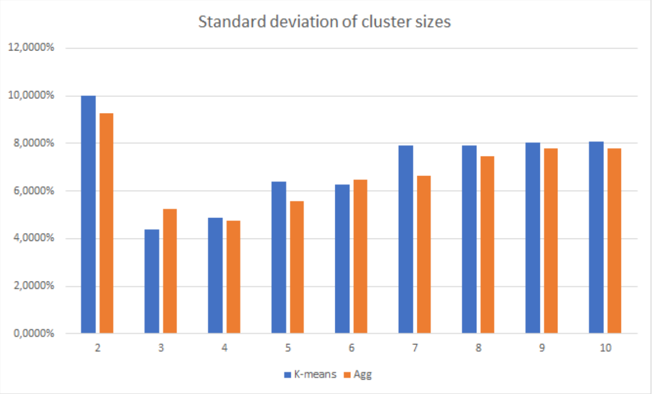

Fig. 2. AIM^2 Standard deviation of cluster sizes

The next characteristic analysed was the size of the largest and smallest cluster. In order to prepare a reasonable dedicated interface variant, the number of customers to whom it will be delivered should be sufficiently large. In this aspect, the DBSCAN method was again found to be useless, as the smallest clusters contained only one client. It turned out that the problem with the number of clusters also occurred with the spectral method. In this case, the largest cluster covered more than 93% of all customers, which calls into question the sense of preparing interface variants for the remaining clusters.

In summary, the results obtained from the experiment allowed to select two of the most promising approaches - agglomerative clustering and K-means method. A detailed analysis was carried out for them, taking into account the different clustering parameters. Further research looked at the impact of clustering parameters on the results obtained. The primary parameter influencing the results was the clustering method, with the additional ones being the fixed number of clusters, pca, init and max_iter.

Firstly, it was checked how the cluster size changes, depending on a fixed number of groups (from 2 to 10), with constant values for the other parameters (pca=0.999, init=k++, max_iter=300).

For the data set analysed, it turned out that the smallest standard deviation of cluster counts was calculated in both methods for 3–4 clusters (Fig. 2). Additionally, it was noted that the agglomerative clustering method yielded a lower standard deviation of cluster sizes in most cases. This means that clusters with less variation can be expected. From the point of

TABLE I
CLUSTERING EFFECTS OF THE METHODS AVAILABLE IN AIM^2

Clustering method	**Clustering time [mins]**	**Number of clusters**	**Largest cluster**	**Smallest cluster**
K-means	4	10* (fixed)	29,7389%	3.4895%
Agglomerative	122	10* (fixed)	29,7302%	2.9438%
DBSCAN	15	282	29.7302%	1 customer
Spectral	188	10* (fixed)	93.6353%	0.1993%

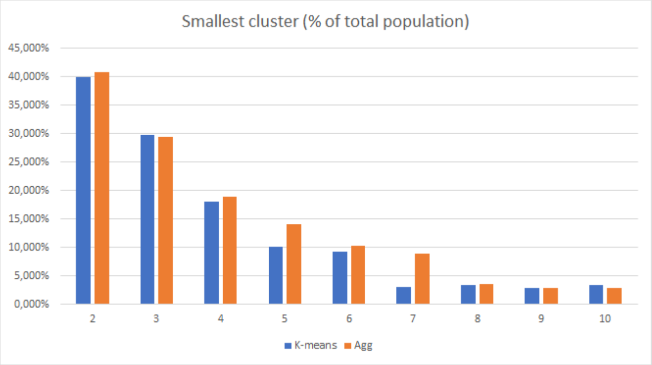

Fig. 3. Smallest clusters

view of variants of the e-commerce interface, agglomerative clustering therefore appears to be slightly better in this aspect. Similar conclusions can be drawn from an analysis of the size of the smallest clusters (Fig. 3). Assuming a cost-efficiency threshold for the development of an interface variant of 5% of the customer population size, it appears that results for more than 6 clusters with the K-means method and for more than 7 clusters with agglomerative clustering should be rejected. The difference does not seem large, but can be significant when choosing the optimal clustering parameters for the delivery of a dedicated e-commerce interface.

When selecting clustering parameters due to business requirements, typical clustering quality indicators cannot be overlooked. Cluster quality analysis for agglomerative clustering and the K-means method, with different numbers of clusters, was carried out for the indices: Silhouette score, entropy, Calinski-Harabasz (CH) score and Davies-Bouldin (DB) score. Given that the first three indicators should be as high as possible and DB as low as possible, it was assumed for simplicity that the analysis would include DB' metric:

$$DB' = \frac{1}{DB} \tag{1}$$

Under this assumption, the goal of clustering optimization is to maximize the value of all quality indicators. For the purpose of the analysis, the values of the indicators were standardized, assuming that the highest value is 100% (Fig. 4).

The results show that there is a very high correlation between the values of quality indicators for both clustering methods. In addition, it can be seen that the values of two indicators increase as the number of clusters increases, and the values of two indicators decrease. The intersection point for the three indicators is a value of 4 clusters. This value can be considered the minimum number of clusters worth generating for a multi-variant e-commerce interface.

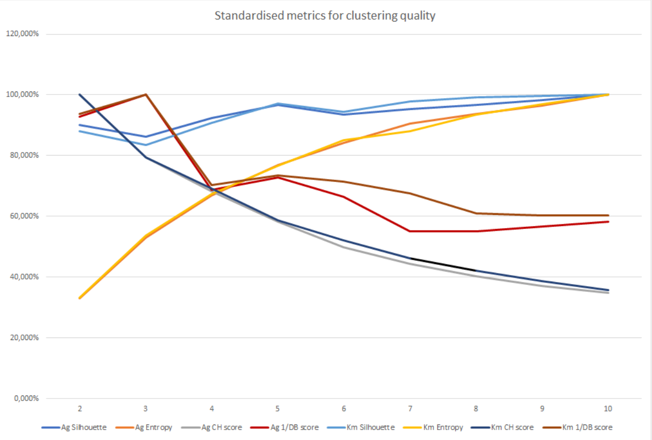

Fig. 4. Standardised metrics for clustering quality

The study showed that the other clustering parameters available in AIM^2 affect the clustering results to a lesser extent.

Increasing the value of max_iter (even 10 times) had no effect on the resulting clusters, so the default value (300) is sufficient.

On the other hand, the option of using the pca parameter to reduce cluster generation time with agglomerative clustering may be interesting. After setting pca=0.95, agglomerative clustering took 13 minutes, almost 10 times faster than calculations with pca=0.99. At the same time, the composition of the clusters has practically not changed, so the reduction in the quality of the input data for clustering should be acceptable.

Slightly more influential is the decision to initialize cluster centers. In the case of random selection of cluster centers, the clusters differ from the selection of the 'k-means++' option. However, the changes applied only to individual clusters, and the standard deviation of cluster sizes was larger each time for the random selection of cluster centers.

Considering the results obtained, for the data set used in the experiment, it could be recommended to use agglomerative clustering with 4 clusters [Fig. 5, 6], due to the smallest variance in cluster size. The number of clusters could be increased (up to a maximum of 7 clusters) if business considerations required it. If cluster calculation time needs to be reduced, the pca parameter could be further changed to 0.95.

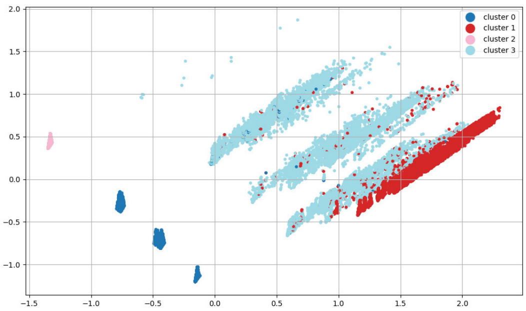

Fig. 5. Scatter plot along the two most significant dimensions produced by the PCA decomposition of the initial dataset

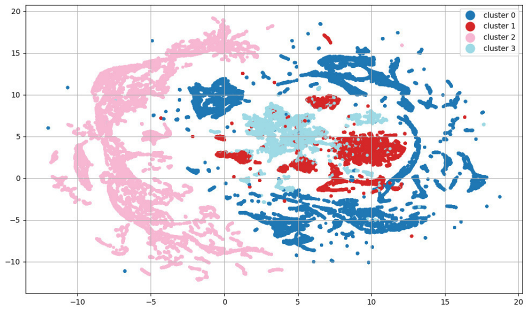

Fig. 6. Scatter plot along the dimensions of low-dimensional representation produced by the UMAP algorithm[14]

V. Conclusion

In this paper, different clusterization methods for multi-variant e-commerce interfaces were reviewed and their performance was compared. Summarizing the results of the study, it can be concluded that after taking into account business requirements and clustering quality metrics, agglomerative clustering for 4-7 clusters or K-means method for 4-6 clusters can be selected for clustering e-commerce customers. In addition, when choosing agglomerative clustering, it is possible to reduce the value of the pca parameter, in order to speed up calculations. The research has identified the most promising clustering methods that can be used to provide specific groups of e-commerce customers with dedicated user interface variants. The results obtained will be used to further develop and validate the AIM^2 platform in future implementations.

Acknowledgment

The datasets used during the study are available from the author upon reasonable request.

References

[1] J. Ah-Pine, ”An Efficient and Effective Generic Agglomerative Hierarchical Clustering Approach”. *Journal of Machine Learning Research* vol 19(1), 2018, pp. 1615-1658.

[2] F. Andriyani, Y. Puspitarani ”Performance Comparison of K-Means and DBScan Algorithms for Text Clustering Product Reviews”, *SinkrOn* vol. 7(3), 2022, pp. 944-949.

[3] L. Ardissono, A. Goy, G. Petrone, M. Segnan ”A multi-agent infrastructure for developing personalized web-based systems” *ACM Transactions on Internet Technology* vol. 5(1), 2005, pp. 47-69.

[4] D. Arthur, S. Vassilvitskii ”k-means++: the advantages of careful seeding”, *SODA '07: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, 2007, pp. 1027-1035.

[5] R.P. Chatterjee, K. Deb, S. Banerjee, A. Das, R. Bag ”Web Mining Using K-Means Clustering and Latest Substring Association Rule for E-Commerce”, *Journal of Mechanics of Continua and Mathematical Sciences* vol. 14(6), 2019, pp. 28-44.

[6] X. Chen, W. Sun, B. Wang, Z. Li, X. Wang, Y. Ye “Spectral Clustering of Customer Transaction Data With a Two-Level Subspace Weighting Method” *IEEE Transactions on Cybernetics* vol. 49, 2019, pp. 3230.

[7] Darwin, R. Purba, M.F. Pasha “Search Query Clustering Comparation On E-Commerce Using K-Means And Adaptive DBSCAN” *3rd International Conference on Mechanical, Electronics, Computer, and Industrial Technology (MECnIT)*, 2020, pp. 207-211.

[8] M. Ester, H.-P. Kriegel, J. Sander, X. Xu ”A density-based algorithm for discovering clusters in large spatial databases with noise”, *KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996, pp. 226-231.

[9] F. Gözükara, S.A. Özel “An Incremental Hierarchical Clustering Based System For Record Linkage In E-Commerce Domain” *The Computer Journal* vol. 66(3), 2021, pp. 581-602.

[10] P.D.Hung, N.T.T. Lien, N.D. Ngoc “Customer Segmentation Using Hierarchical Agglomerative Clustering” *ICISS '19: Proceedings of the 2nd International Conference on Information Science and Systems*, 2019, pp. 33-37.

[11] P. Jiang, Y. Zhu, Y. Zhang, Q. Yuan “Life-stage Prediction for Product Recommendation in E-commerce” *KDD '15: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 1879-1888.

[12] A. Kobsa, “Personalized hypermedia and international privacy” *Communications of the ACM* vol. 45, 2002, pp. 64-67.

[13] R.V.S. Kumar, S.S. Rao, P. Srinivasrao “An Efficient Clustering Approach using DBSCAN” *Helix* vol. 8(3), 2018, pp. 3399-2305.

[14] L. McInnes, J. Healy, J. Melville “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction” *ArXiv e-prints*, 2020, https://arxiv.org/pdf/1802.03426

[15] M. Montaner, B. López, J. L. de la Rosa “A Taxonomy of Recommender Agents on the Internet” *Artificial Intelligence Review* vol. 19, 2003, pp. 285-330.

[16] E. Patel, D.S. Kushwaha “Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model” *Artificial Procedia Computer Science* vol. 171, 2020, pp. 158-167.

[17] K. Tabianan, S. Velu, V. Ravi ”K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data”, *Sustainability* vol. 14(12), 2022, pp. 1-15.

[18] E. Triandini, F.A. Hermawati, I.K.P. Suniantara ”Hierarchical Clustering for Functionalities E-Commerce Adoption”, *Jurnal Ilmiah KURSOR* vol. 10(3), 2020, pp. 111-118.

[19] Y. Yang, J. Jiang, H. Wang ”Application of E-Commerce Sites Evaluation Based on Factor Analysis and Improved DBSCAN Algorithm”, *International Conference on Management of e-Commerce and e-Government*, 2018, pp. 33-38.

[20] L. Wang, Y. Jing ”Collocating Recommendation Method for E-Commerce Based on Fuzzy C-Means Clustering Algorithm”, *Journal of Mathematics* vol. 2022, 2022, pp. 1-11.

[21] Z. Wu, L. Jin, J. Zhao, L. Jing, L. Chen ”Research on Segmenting E-Commerce Customer through an Improved K-Medoids Clustering Algorithm”, *Computational Intelligence and Neuroscience* vol. 2022, 2022, pp. 1-10.

[22] B. Zhang, L. Wang, Y. Li ”Precision Marketing Method of E-Commerce Platform Based on Clustering Algorithm”, *Complexity* vol. 2021, 2021, pp. 1-10.

[23] Y. Zhang, M. Li, S. Wang, S. Dai, L. Luo, E. Zhu, H. Xu, X, Zhu, C. Yao, H. Zhou ”Gaussian Mixture Model Clustering with Incomplete Data”, *ACM Transactions on Multimedia Computing, Communications, and Applications* vol. 17(1s), 2021, pp. 1-14.