Customer Online Shopping Feature Extraction Based on Data Mining Algorithm

Chia-Chih Chen[1] and Thai-Ha Lin[2]

Chung Yuan Christian University, Taoyuan City, Taiwan
chiachih53216@gmail.com

Abstract

Driven by the climax of the Internet, online shopping has brought people into a new shopping era and has also brought new impacts to enterprises. To improve the market competitiveness of enterprises, enterprises need to continuously mine customer behavior information. In the mining process, due to the high amount of customer behavior characteristics, the existing behavior mining processing has problems such as low acceleration and high error rate. Feature extraction customer behavior mining algorithm, this algorithm estimates the non-customer behavior and customer behavior in online shopping, iterates many times until convergence, and obtains the best mining result corresponding to the regression line and variance feature parameters, and completes the customer behavior mining. The simulation test proves that the proposed algorithm can improve the precision and recall rate, ensure the reliability and stability of customer behavior mining, and has certain use-value in practical applications.

Keywords: Data mining, Customer behavior, Accuracy

1. Introduction

With the popularization of the Internet of Things, it is precisely because online payment is used to meet the various requirements of buyers and retailers that online shopping has obtained opportunities for rapid development and has gradually become an important new shopping model [1]. As people's living standards continue to rise and the pace of life continues to accelerate, retailers will serve consumers on the condition of saving time and energy [2]. In this highly competitive society, to improve the market competitiveness of enterprises, it is necessary to collect consumer information and dig out the characteristics of customers' online shopping behaviors to determine marketing strategies that meet the market conditions and enhance the competitiveness of enterprises [3]. In the process of mining customer behavior information, due to complex customer behavior characteristics and high redundancy, the existing behavior mining processing technology has problems such as low mining acceleration, high error, and high energy consumption. In response to the above problems, it is necessary to explore an effective customer behavior mining algorithm to ensure the reliability and practicability of behavior mining [4].

Literature [5] proposed a customer behavior mining algorithm based on the FP-Growth algorithm, using redundant dictionaries to reduce the redundancy of customer behavior mining results; giving different support weights according to the length of the behavior, and at the same time improving the FP-Growth algorithm, According to the degree of independent support and

the confidence of behavior features, feature extraction is performed on the features to be selected, which realizes customer behavior mining. Literature [6] proposed a customer behavior mining algorithm based on neural network adaptive search. The algorithm established a neural network training model for customer behavior feature mining and improved BP neural network training to cluster and extract customer behavior features. The adaptive search iterative method improves the customer behavior mining algorithm, enhances the clustering and feature extraction results of customer behavior mining, and realizes customer behavior mining. Literature [7] proposed a customer behavior mining algorithm based on the improved Aprion algorithm. Under the frequent pattern tree, using the service-oriented architecture SOA model to construct a frequent pattern tree for mining frequent items of customer behavior. At the same time, the behavior information is fused and preprocessed, and the concept lattice of behavior item set association rules is constructed to extract customer behavior characteristics and realize customer behavior mining.

In summary, although the above methods have a certain degree of efficiency in a period, with the continuous expansion of the scale of online shopping customers, the amount of customer behavior characteristics is relatively high. According to these redundant items of online shopping characteristics, customers are Behavior mining has problems such as speedup, precision, and recall. In response to the above problems, a customer behavior mining algorithm based on online shopping feature extraction is proposed, and simulation tests verify that the proposed algorithm can effectively solve the above problems.

2. Customer behavior mining algorithm for online shopping feature extraction

2.1. Preprocessing of customer behavior characteristics

Conduct behavior analysis based on consumers' online shopping characteristics, formally express customer's online shopping behavior data sets, form the information expression of customer transaction data collection transaction statistical information, obtain customer online shopping behavior rules, and use the guidelines to extract customer online shopping behavior characteristics.

Hypothesis $K = (XC, I, WXC)$ represents the customer data set...$XC =| xc_1, xc_2, \cdots xc_{|XC|} |$ represents a collection of customers.$| XC |$ represents the number in the set.$I = \{i_x | x = 1, 2, \cdots, n\}$ represents the customer attribute set. $WXC = \{Wi_x | x = 1 \quad 2, \cdots, n\}$ ， Wi_x represents the value range of customer attribute i_x [8]，The projection feature of customer $xc \in XCi_x$ on i_x is$XC \to Wi_x, i_x(xc) \in Wi_x$。

Hypothesis $H = (R, XC \cup P, WP)$ represents a collection of online shopping transactions. $R = \{r_1, r_1, \cdots r_{1R1}\}$ represents a collection of online shopping transactions,$|R|$ represents the number of vectors in the set.$P = \{p_x | x = 1, 2, \cdots, n\}$ represents the collection of purchased items; $XC \neq \Phi$ ， $P \neq \Phi, XC \cap P = \Phi; WP = \{W_{p_x} | x = 1, 2, \cdots, n\}$ ， W_{p_x} represents the shopping volume of the purchased item p_x, and the behavior r $\in$ R is projected on p_x as p_x: R $\to W p_x$,$p_x(r) \in W p_x$。

Hypothesis $HD = (R, X, W)$ represents a collection of customer online shopping transaction data. Online shopping feature set$X = XC \cup I \cup P, W = WXC \cup WP_\circ$。 When satisfied $H \subseteq X, C \subseteq X, H \cap C = \Phi$， H $\cup$ C = X， H represents the set of conditional attributes, and C represents the set of behavioral attributes [9-10].

Assume $HD' = (R, H \cup C, W)$ represents a subset of customer online shopping feature vectors, set $H \subseteq X, C \subseteq X$, and $H \cap C = \Phi$, The expression of online shopping transaction rule O is as follows

$$\begin{cases} \langle h, w_h \rangle | h \in H, w_h \in Wh, Wh \in W| \Rightarrow \\ \{\langle c, w_c \rangle c \in C, w_c \in W_c, W_c \in W\}(l) \end{cases} \tag{1}$$

In the formula: l belongs to the uncertainty factor, and its expression is as follows

$$l = \frac{|\{r \mid r \in R, h \in H, h(r) = w_h, \langle h, w_h \rangle \in 0\}|}{|\{r \mid r \in R, c \in C, h(r) = w_c, \langle h, w_c \rangle \in 0\}|} \tag{2}$$

In the formula, $l \in [0,1]$ is the feature extraction ability of transaction attributes relying on conditional attributes. The set of characteristic rules for online shopping behavior is as follows:

$$OSF = \{ Path \mid Path = O \} \tag{3}$$

The division of customer online shopping transaction data set HD is represented by $CxwR(R) = |R_x|R_x \subseteq R|$, and the conditions met are as follows:

$$\forall R_1, R_2 \in CxwR(R), R_1 \cap R_2 = \Phi$$
$$R = \cup_{\forall R_k \in Cun(R)} R_x \tag{4}$$

In attribute set A, $A \subseteq X, R/A$ represents a division related to HD, and the conditions satisfied are as follows

$$\forall R_x \in \frac{R}{A}, \forall v, w \in R_x,$$
$$\forall a \in A, a(v) = a(w) \tag{5}$$

Where, $R_x \in R/A$ represents an equivalent category related to A, R/A represents a family of equivalence related to A. If $o_x = o(R_x, A) = |R_x|/|R|$ represents the probability of the relevant equivalent category. According to the rules of online shopping transactions, use the following formula to give the probability distribution of the attribute set A related to the customer data set, and extract the characteristics of customer online shopping behavior

$$O(R, A) = \{|o(R_x, A), |\forall R_x \in \frac{R}{A}\} \tag{6}$$

In the process of extracting the above-mentioned online shopping features, there are often a large number of high-dimensional features and noises. These high-dimensional features and noise not only increase the time of customer behavior mining but also affect the accuracy of online shopping feature extraction to a certain extent. Aiming at this characteristic, a more effective feature extraction regularization estimation method for high-dimensional online shopping data is proposed.

2.2. Customer behavior mining algorithm based on feature extraction of online shopping

Based on the above-mentioned customer online shopping behavior feature extraction results, a feature extraction regularization estimation method that is more effective for high-dimensional online shopping features is proposed. The current feature extraction regularization estimation mainly focuses on regression coefficients. The online shopping feature set obtained in Section 2.1 is high-dimensional Data, in addition to the estimation of the regression coefficient α, μ^2 is also an important parameter. After minimizing it, the regularization

estimation method is used to construct the online shopping feature μ^2 minimization estimation model, and its expression is as follows

$$\grave{\alpha}, \grave{\mu}^2 = \underset{\alpha, \mu^2}{argmin}\left\{-\ln V(\alpha, \mu^2; R_1, \cdots, R_n) +\gamma \sum_{i=1}^{n} \grave{\omega}_i |\alpha|\right\} \qquad (7)$$

Where,$\ln V(\alpha, \mu^2; R_1, \cdots, R_n)$ represents the log-likelihood function in the online shopping feature minimization estimation model [9]; V represents the likelihood function; α represents the regression coefficient; μ^2 represents the variance of the characteristic error; γ represents the penalty parameter; $\widetilde{\omega}_i (i = 1, \cdots, n)$ represents the least-squares coefficient. If you want to consider the characteristic variance parameter of online shopping Estimated to be affected by the regression coefficient α, the equation (8) needs to be transformed into the following form

$$\grave{\alpha}, \grave{\mu}^2 = \underset{a, \mu^2}{argmin}\left\{\frac{1}{2}\ln \mu^2 + \frac{1}{2\mu^2}\sum_{j=1}^{m} \left(R_j - A_j^o \alpha\right)^2 + \gamma \sum_{i=1}^{n} \grave{\omega}_i |\alpha|\right\} \qquad (8)$$

In formula (8), the penalty used is adaptive, and only the regression coefficient is penalized, and α and μ^2 are minimized. The penalty parameter and the estimated value of $\grave{\alpha}$ can directly affect the estimation of the variance characteristic parameter $\grave{\mu}^2$.

When the scale of the attribute variable changes, both $\grave{\alpha}$ and $\grave{\mu}^2$ have scale invariance. The objective function in equation (8) is non-convex. Minimization makes feature extraction more complicated and unstable. If you want to avoid these two problems, the expressions that need to be converted from (8) are as follows

$$\grave{\alpha}, \grave{\mu}^2 = \underset{\alpha, \mu^2}{argmin}\left\{\frac{1}{2}\ln \mu^2 + \frac{1}{2\mu^2}\sum_{j=1}^{m} \left(R_j - A_j^o \alpha\right)^2 + \gamma \sum_{i=1}^{n} \grave{\omega}_i \frac{|\alpha|}{\mu}\right\} \qquad (9)$$

The converted formula (9) can solve the invariance of the scale. If the parameters are converted to $Y_i = \alpha_i/\mu, \delta = \mu$, the converted expression of formula (9) is as follows

$$\grave{Y}, \grave{\delta}^2 = \underset{a, \mu^2}{argmin}\left\{\frac{1}{2}\ln \delta + \frac{1}{2\mu^2}\sum_{j=1}^{m} \left(\delta R_j - A_j^o Y\right)^2 + \gamma \sum_{i=1}^{n} \grave{\omega}_i |Y_i|\right\} \qquad (10)$$

In the formula, $Y = (Y_1, \cdots, Y_n)$, then the objective function is converted to a convex function.

2.3. Customer behavior mining algorithm based on online shopping feature extraction

Taking the regularized estimated value of online shopping feature extraction in section 2.2 as input and combining the coordinate algorithm and KKT condition, calculate the new customer online shopping feature extraction regularized estimated value, iterate multiple times until convergence, and output the best result corresponding to the regression line and the variance feature parameter. Complete customer behavior mining.

The online shopping feature extraction algorithm combines the coordinate algorithm with the KKT condition to obtain the regularized estimate $\grave{Y}, \grave{\delta}$, and further completes the customer behavior mining based on the estimation results.

Using the KTT condition, the conditions satisfied by $\grave{Y}$ and $\grave{\delta}$ in equation (10) are as follows

When $\grave{Y}_i = 0, \dfrac{\left|-\grave{\delta} N_i^o M + N_i^o N \grave{Y}\right|}{\grave{\omega}} \leqslant \gamma_0$

When $\dot{Y}_i \neq 0, -\hat{\delta} N_i^o M + N_i^o N\dot{Y} + \gamma \dot{\omega}_i \mathrm{sgn}(\dot{Y}_i) = 0$

Regularized estimation of customer behavior

$$\hat{\delta} = \frac{\langle M, N\dot{Y} \rangle + \sqrt{\langle M, NY \rangle^2 + 4 \parallel M \parallel^2}}{2 \parallel M \parallel^2} \tag{11}$$

Where, $M = (R_1, \cdots, R_m)$ represents m-dimensional vector; $N = (A_1, \cdots, A_m)$ represents $m \times n$-dimensional matrix; N_i represents the i-th column of N; $\mathrm{sgn}(\cdot)$ represents symbolic function; $\langle \cdot \cdot \rangle$ represents behavior; $\parallel \cdot \parallel^2$ represents the Euclidean distance of the behavior vector.

According to the above KKT condition, obtain the result of $\hat{\delta}$, make $\gamma_t = \gamma_{\max} \varphi^{k/100} (\kappa = 1, \cdots, 100)$, Assume $\varphi = 0.01, \kappa = 0$ represents the spatial dimension, the calculation formula for the maximum value of the adaptive penalty function is as follows

$$\gamma_{max} = max \left| \frac{\langle N_i, M \rangle}{\parallel M \parallel} \right| \tag{12}$$

Order $\kappa = \kappa + 1$, then $\gamma = \gamma_\kappa$, use the coordinate algorithm to obtain the original customer behavior estimate of $\dot{\alpha}^{(0)}$, order $b = 0, \dot{Y}^6 = \dot{\alpha}^{(b)}/\dot{\mu}^{(b)}, \hat{\delta}^b = 1/\dot{\mu}^{(b)}$, use variance μ^2 to estimate the original customer behavior, as follows

$$\delta^{(0)2} = \frac{1}{m} \sum_{j=1}^{m} \left(R_j - A_i^o \dot{\alpha}^{(0)} \right)^2 \tag{13}$$

Order $\widetilde{\omega}_i = \min\{1/|\dot{Y}_i^b|, 10^6\}, i = 1, \cdots, n$, the new calculation formula for customer behavior estimation is as follows

$$\hat{\delta}^{(b+1)} = \left[\langle M, N\dot{Y}^{(b)} \rangle + \sqrt{\langle M, N\dot{Y}^{(b)} \rangle^2 + 4 \parallel \frac{M \parallel^2}{(2 \parallel M \parallel^2)}} \right] \tag{14}$$

$$\dot{Y}_i^{(b+1)} = \begin{cases} 0, |F_i| \leqslant \widetilde{\omega}_i \gamma_\kappa \\ \dfrac{(\gamma_\kappa \dot{\omega}_i - F_i)}{\|N_i\|^2}, F_i > \gamma_\kappa \dot{\omega}_i \\ -\dfrac{(\gamma_\kappa \widetilde{\omega}_i + F)}{\|N_i\|^2}, F_i < -\gamma_\kappa \dot{\omega}_i \end{cases} \tag{15}$$

Where, $F_i = -\hat{\delta}^{(6+1)} A_i M + \sum_{i=1, l \neq j}^{m} \dot{Y}_t^{(b+1)} N_i^o N_t$ represents the estimated number of customer behaviors.

Repeat steps 2) and 3) many times until convergence.

When κ<100, go back to step 1), otherwise the extraction ends.

The estimated value of customer behavior of μ^2 and α obtained by two algorithms, customer behavior data

$$\gamma_{cbm} = \mathrm{m} \ \{O(\gamma_\kappa), \kappa = 1, \cdots, 100\} \tag{16}$$

3. Simulation and analysis

To prove the effectiveness of the customer behavior mining algorithm based on online shopping feature extraction, the proposed algorithm is compared with the literature [5]

algorithm, the literature [6] algorithm, and the literature [7] algorithm. The total rate and the comprehensive value of these two items are used as evaluation indicators for experimental testing.

The simulation is performed on the CPU Intel Pentium; the main frequency is 3.4GHz, the memory is 8GB, the operating system is Windows7 Pro operating system, and the customer behavior mining simulation is performed in the Matlab environment.

1) Mining speedup

Speedup is the ratio of time consumed in a single processor system and a parallel processor system, which can reflect the energy consumption and speed of customer behavior mining in online shopping. If the mining acceleration is relatively low, it means that the speed of customer behavior feature mining is slower and the energy consumption is high. If the mining acceleration is relatively high, it means that the speed of customer behavior mining is faster, the energy consumption is low, and the mining effect is better.

In the simulation, the selected speedup ratio values are 800, 1200, 1600, 1800, 2200, and 2600, as shown in [Figure 1]. The comparison results of the mining speedup ratios of different mining algorithms are shown in [Figure 1].
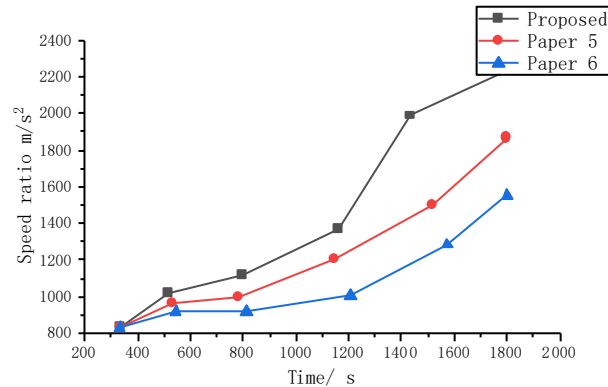

Figure 1. Mining acceleration ratio under different mining algorithms

It can be seen from [Figure 1] that three different algorithms are used to mine customer behavior. The speedup ratio of the three algorithms mining increases with time. When the time is 400s ~800s, the mining speedup ratio of the three algorithms is always Below 1200m/s2; when the time is about 1000s, the mining acceleration ratios of the three algorithms begin to rise rapidly. When the time is about 1400s, the mining acceleration ratio has reached about 1400m/s2 using the algorithm of literature [6], using the algorithm of literature [5] The mining speedup ratio reached about 1650m/s2, and the mining speedup ratio of the customer behavior mining algorithm based on online shopping feature extraction reached about 1980m/s2; in contrast, the mining speedup comparison between the algorithm in literature [5] and the algorithm in literature [6] The main reason is that in the data mining process of these two algorithms, the single-processor system and the parallel processor system consume a long time. As a result, the two mining algorithms cannot quickly find the best customers in a short mining time. Behavior information, and lead to high mining acceleration and poor mining quality, while the proposed customer behavior mining algorithm based on online shopping feature extraction has a relatively high mining acceleration, indicating that the proposed method can improve mining efficiency and enhance mining in a short period Speedup ratio.

[Figure 2] is a simulation test result comparing the error rate of the customer behavior mining algorithm based on online shopping feature extraction and the algorithm mining error rate of the literature [7] under different amounts of customer behavior in online shopping.
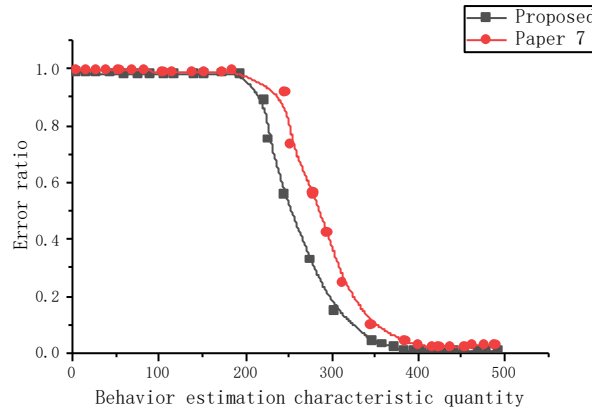

Figure 2 Comparison of the error rate of the two mining algorithms

It can be seen from [Figure 2] that the mining error rate decreases with the increase of customer behavior estimated feature quantity. When the customer behavior estimated feature, quantity is between 0 and 200, the mining error rate of the two mining algorithms is the same. The estimated feature quantity of customer behavior starts from 200, and the error rate of the two gradually decreases. When the estimated feature quantity of customer behavior is between 200 and 400, the error rate of the algorithm using literature [7] is significantly higher than that of the customer behavior mining algorithm based on online shopping feature extraction. In this process, the estimated feature quantity of customer behavior continues to increase. When using the literature [7] algorithm mining, it is not possible to accurately estimate the feature quantity of the selection behavior, and most of the obtained characteristics are not the customer behavior estimation feature, which leads to a higher error rate of the mining results. However, the proposed mining algorithm is used to estimate the characteristic quantities of customer behavior, and after multiple iterations, accurate mining results can be obtained, and the error rate is low. After the amount of customer behavior data is 400-500, the error rates of the two mining algorithms almost intersect. If you look carefully at the image, you can see that the error rate of the proposed mining algorithm is still lower than that of the algorithm in the literature [7]. This proves that the customer behavior mining algorithm based on online shopping feature extraction can reduce the error rate and has a certain degree of reliability under the condition that the quality of the mining results can be guaranteed.

2) Performance evaluation

Mining precision and recall rates can reflect the reliability, usability, and robustness of customer behavior data mining in online shopping. If the mining accuracy rate is lower, it means that the reliability, availability, and robustness of customer behavior data mining is lower. If the mining accuracy rate is higher, it means that the reliability, availability, and robustness of customer behavior data mining is higher. high.

The evaluation indexes of the experimental performance are recall rate, precision rate, and the comprehensive value of these two items. The test results of the comparison test of four different mining algorithms are shown in [Figure 2] and [Table 1]. Suppose G represents the number of behavioral features mined by algorithm; Y represents the number of behavioral

features mined by algorithm; Q represents the number of behavioral features that have not been extracted, the recall rate, precision rate, and the formula of comprehensive value measurement indicators as follows

$$L = \frac{G}{G + Y} \tag{17}$$

$$D = \frac{G}{G + Q} \tag{18}$$

$$J - s\,c\,o\,r\,e = \frac{2LD}{L + D} \tag{19}$$

In [Table 1], AL represents the mining algorithm; AM represents the proposed algorithm; Fa represents the document [5] algorithm; Sa represents the document [6] algorithm; Se represents the document [7] algorithm; Pr represents the precision rate; Re represents the recall rate; Cv represents the integrated value.

Table 1. Comparison of different evaluation indicators under different mining algorithms

AL	Pr/ %	Re/ %	Cv/ %
AM	99.17	95.61	97.39
Fa	95.46	86.7	91.08
Sa	90.7	92.1	91.4
Se	98.99	91.7	95.3

According to the above [Table 1], it can be seen that four mining algorithms are used to mine customer behavior characteristics. In the accuracy comparison, the accuracy of the algorithm in literature [7] is higher than the algorithm in literature [5] and the algorithm in literature [6], literature [5] The precision of algorithm mining is the lowest, and the precision of the proposed algorithm is higher than that of the algorithm in literature [7]; in the comparison of recall, the precision of the algorithm in literature [6] is higher than that of the algorithm in literature [5]. Literature [7] has a high algorithm, and the accuracy of the algorithm in literature [5] is the lowest. The accuracy of the proposed algorithm is higher than that of the algorithm in the literature [6], indicating that the behavioral features extracted by the proposed algorithm are relatively complete. Few behavioral characteristics ensure the final precision rate; in the comprehensive value comparison, the comprehensive value of the proposed mining algorithm is higher than the other three mining algorithms, which proves that the proposed mining algorithm can ensure the integrity of the mining behavioral characteristics and stability. From the above, the proposed customer behavior mining algorithm based on online shopping feature extraction can effectively improve the mining speedup ratio, reduce the error rate, ensure the precision and recall rate, and make the mining results more reliable and usable.

4. Conclusion

Exploring an effective customer behavior mining algorithm can increase the mining speed-up ratio, reduce the error rate, and ensure the reliability and practicability of behavior mining. To solve the problems of high error and low acceleration in customer behavior mining processing, a customer behavior mining algorithm based on online shopping feature extraction is proposed. The algorithm is based on the ability to extract the characteristics of customer behavior condition attributes based on transaction attributes to obtain online shopping feature rules. Using the regularization estimation method, the regression line estimated value and the

variance characteristic parameter estimated value are minimized, and the coordinate algorithm and the KKT condition are combined to obtain the regression line and the variance characteristic parameter. The best mining results corresponding to the regression line and variance characteristic parameters were mined using the online shopping criteria, and the customer behavior mining was completed. The simulation results show that the proposed algorithm can effectively increase the mining speed-up ratio, reduce the error rate, and improve the recall and precision rates.

References

[1]   J. F. Henriques, "High-speed tracking with Kernelized correlation filters," IEEE Transactions on Pattern Analysis & Machine Intelligence, vol.37, no.3, pp.583-596, **(2015)**

[2]   M. Danelljan, "Adaptive color attributes for real-time visual tracking," IEEE Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, pp.1090-1097, **(2014)**

[3]   M. A. Nartea and J. Jamice, "Exploring the characteristics of millennials in online buying behavior, **(2019)**

[4]   L. Bertinetto, "Staple: Complementary learners for real-time tracking," vol.38, no.2, pp.1401-1409, **(2015)**

[5]   M. Mueller, N. Smith, and B. Ghanem, "Context-aware correlation filter tracking," IEEE Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, pp.1387-1395, **(2017)**

[6]   M. Wang, Y. Liu, and Z. Huang, "Large margin object tracking with circulant feature maps," Computer Vision and Pattern Recognition (cs. CV), **(2017)**

[7]   A. C. Silva and C. Borges, "An improved heuristic-based genetic algorithm for bin packing problem," 2019 8th Brazilian Conference on Intelligent Systems (BRACIS), **(2019)**

[8]   C. Ma, "Hierarchical convolutional features for visual tracking," IEEE International Conference on Computer Vision, IEEE, pp.3074-3082, **(2016)**

[9]   P. Duarte, C. E. S. Susana, and M. B. Ferreira, "How convenient is it? Delivering online shopping convenience to enhance customer satisfaction and encourage e-WOM," Journal of Retailing and Consumer Services, vol.44, pp.161-169, Sept. **(2018)**

[10] S. K. Ariffin, T. Mohan, and Y. N. Goh, "Influence of consumers' perceived risk on consumers' online purchase intention," Journal of Research in Interactive Marketing, vol.12, **(2018)**

This page is empty by intention.

Chia-Chih Chen and Thai-Ha Lin