
Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 5, pp. 3515–3522, DOI 10.12694/scpe.v25i5.3041

E-COMMERCE DATA MINING ANALYSIS BASED ON USER PREFERENCES AND ASSOCIATION RULES

YUN ZHANG*

Abstract. With the development of network technology, online shopping is becoming more and more convenient. But the increasing number of products also makes it difficult for consumers to make the right decision. When there is no apparent market demand, how to recommend products with commercial potential to customers has become an urgent problem for businesses to solve. This paper proposes e-commerce product recommendation based on user preference and association rule algorithm aiming at the problems existing in e-commerce product recommendation. Firstly, this paper constructs a user interest modeling method. Through analyzing users' interests and preferences, to provide users with timely and accurate personalized services. Then, the FP_Growth algorithm is optimized and improved. A more effective CTE-MARM algorithm is designed, and an association rules database based on user benefit items is constructed and analyzed jointly. Analyze products with strong correlations. According to consumers' interest levels, TOP-N is the best product choice. Experiments show that the algorithm has higher prediction accuracy. The research results of this project can not only improve enterprises' ability to analyse data and provide data support for enterprises to carry out effective marketing management.

Key words: Data mining; Association rules; User preference; Electronic commerce; CTE-MARM algorithm

1. Introduction. In modern society, with the rapid development of science and technology, human beings are also faced with the problem of "excess" and the convenience of obtaining science and technology. This phenomenon is no exception in e-commerce. Making customers satisfied with products is a significant issue for e-commerce enterprises. This is how the product recommendation technology in e-commerce emerged [1]. It applies the method of data mining to the actual consumer behavior scenario. Possible business opportunities can be predicted through the correlation mining of historical data. This saves users time finding items they like and increases sales and customer loyalty.

In the same frame, it is of great theoretical value and practical significance to study three different types of customer evaluation: customer perceived value, customer satisfaction, and customer purchase intention. Previous studies have shown that perceived value is the pre-variable of consumer satisfaction. Risk perception, individual innovation, social impact, and perception of usefulness directly and significantly impact user intent. Perceived ease of use and social influence directly impact user availability but cannot play an indirect role through perceived usefulness. Many psychological barriers of consumers to e-commerce, especially the security and reputation problems faced by e-commerce users, will significantly impact consumers' purchase intentions and cause significant obstacles to the rapid development of e-commerce. After a questionnaire survey and analysis of e-commerce users, some researchers found that social factors have the most apparent positive effect on e-commerce intentions, while they have no noticeable effect on e-commerce expected practicability and risk perception. Mobile phone payment intention positively affects users' use, but convenience has no noticeable promoting effect [2]. Some scholars use the integrated technology acceptance model to study electronic transaction intention from six perspectives: perception of usefulness, perception of ease of use, perception of risk, trust, trust and evaluation, and attitude of use. Relevant studies mainly focus on the perception and evaluation of consumer behavior but lack systematic collection and analysis of mobile payment users' characteristics and future consumer behavior [3]. This makes it a complex problem to accurately position the market of its target users in the initial stage of the development of e-commerce.

*Zhejiang Yuexiu University, Shaoxing, 312000, China (460465456@qq.com)

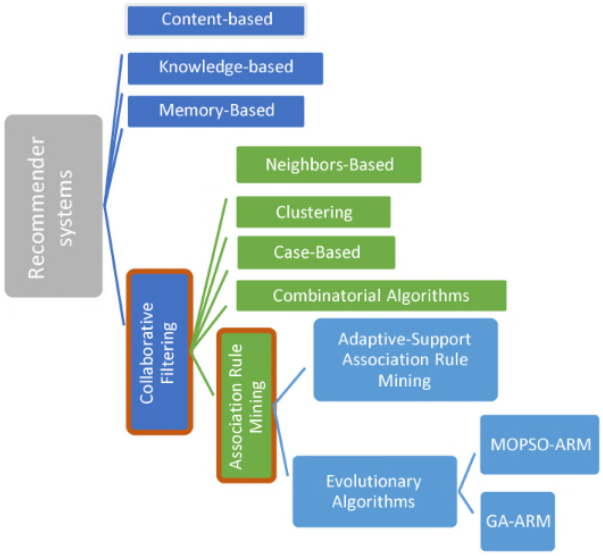

Fig. 2.1: *Association rules model flow.*

2. Demand analysis.

2.1. Functional Requirements. The key to an e-commerce product recommendation system is to analyze the interests and preferences of users by collecting their shopping habits and shopping records. Then, explore and predict the potential shopping opportunities [4]. The most important thing is to analyze users' personalization and real-time performance. The system implements the following essential functions.

1. Data collection: extracted relevant records and operational data.
2. Preprocess and eliminate unnecessary data to ensure data integrity.
3. The user's interest model is established and classified.
4. Establish the relevant rule database of data mining.
5. Product recommendation for products that are attractive to users.

2.2. Key Parameters.

1. Credibility is introduced to reduce the phenomenon of "rule explosion" and improve the search accuracy;
2. The prescription describes the change in consumption behavior in real life. The closer to the current consumption behavior, the more it can reflect the current demand preferences.
3. The purchase, browsing, collection, rating, and comments of e-commerce consumers can reflect consumers' interests [5]. Users have different levels of interest in different business activities. Use the triplet method to sort the required items in order IR_{ij}. The "user-benefit item chain" comprises the N items with the highest information value by the TOP-N method. The CTE-MARM method is proposed for data association [6]. When good $_i$ in the list of customer-benefit items has a strong correlation with good_k, good_k is added to the tripartite group. The model flow diagram is shown in Figure 2.1.

3. Association rule algorithm model design. This topic focuses on how to quickly find the user group with specific characteristics in the process of commercial marketing of e-commerce enterprises. Discover the characteristics of the target user group. This user group has more complex characteristics [7]. The value of the variable is random and difficult to transform. Each attribute is unrelated, so it is challenging to meet the needs of conventional statistical methods using standard multiple regression analysis, structural equation model, technology acceptance model, technology acceptance model, and integrated technology acceptance model. It is challenging to realize the rapid positioning of specific user groups. In the practical problems, from many incomplete, noisy, fuzzy, random, and other practical problems, we extract the information and knowledge that

people do not know in advance but have potential value [8]. Then, the information in these eigenvalues can be obtained by mining the association rules.

3.1. Principles of Association Rules. Association rules were proposed by R. Agrawal in 1993 to describe the internal relationship between various items in a database. It is one of the essential research contents in data mining [9]. Association rules can directly represent the relationship between a collection of items in a data set. These associations are not based on a specific distribution or depend on a specific pattern. It depends only on the probability of occurrence of the set of items in a pattern. Suppose $Q = \{q_1, q_2, \mathrm{L}, q_m\}$ is the set of all entries. H represents the transaction database and T represents project subset ($T \subseteq Q$). Each transaction has its unique transaction identifier $TID.B$ is the set of entries. Transaction record T includes item B. Its necessary and sufficient condition is that if the entry B contains k items, it is called the set of k items. The proportion of the number of items B in the transaction database H is called the level of support for the item set. If the support level of an item set exceeds the minimum support threshold set by the user, it is called a frequent item set. The law of association is the logical implication of class $U \Rightarrow V$. Where $U \subset Q, V \subset Q$, and $U\mathrm{I}V = \varnothing$. If transaction database contains $s\%$ transactions and includes $U\mathrm{U}V$, then the support of trading system $U \Rightarrow V$ is $s\%$. The absolute amount of support is the possible number. If $\sup\mathrm{port}(U)$ is used to represent the degree of support for item U, then $\sup\mathrm{port}(U\mathrm{U}V)/\sup\mathrm{port}(U)$ can be used to represent the degree of confidence of the rule, which is the conditional probability $P(V \mid U)$.

$$\sup\mathrm{port}(U \Rightarrow V) = P(U\mathrm{U}V)$$
$$\mathrm{conf}\,i\,\mathrm{dence}(U \Rightarrow V) = P(V \mid U)$$

The related rules conforming to the minimum support and threshold are called strict rules [10]. The maximum is $0 - 100\%$. That is, whether the item on the right will be selected when the item on the left is purchased or whether the item on the right will be selected in any situation. In the process of selecting target customers, the size of the revenue is the most important. The greater the revenue value, the greater the customer demand for the service. This criterion is the same as the selection criteria for other data exploration modes [11]. To measure how much this criterion improves the prediction accuracy by comparing it with the "original" criterion.

3.2. Evaluate the role of the "promotion" attribute group in association rules. Assume that the mobile user base is constant. It grows at a constant rate over some time. But compared to the number of existing mobile phone users, the number of new mobile phone users is still relatively small [12]. For this reason, the number of users will grow in a particular proportion over some time, so at a certain point in time, the number of mobile phone users can be limited to:

$$\text{total users } = IE$$

I is the invariant growth factor. E is the number of mobile phone users at a given point in time. In association rule $[D, Z, S], D$ is support, Z is credibility, and S is revenue. The number of mobile phone users $\times D$ can be considered to be able to support a certain number of mobile phone users. Under the promotion effect of the previous part, the mobile user number H can promote the latter part. The number of mobile subscribers $\times D \times Z \times S$ can be understood as selling mobile phones to a specific group [13]. The number of mobile users that can make subsequent things happen. Using the number of mobile phones $\times D \times Z \times S$, the "promotion" effect of the former term on the latter term can be evaluated, and the "promotion" effect is more prominent when the effect value is more significant. Since the number of mobile phone users IE is constant at a certain point in time, a "boost" factor is introduced here:

$$\phi = D \times Z \times S$$

D stands for support, Z for credibility, and S for revenue. The "boost" factor is used to measure the strength of the "boost" connection between the first product and the later product [14]. With the increase of ϕ, this "promoting" effect gradually increases. It is targeted at this audience characteristic for marketing publicity, which will receive a good effect.

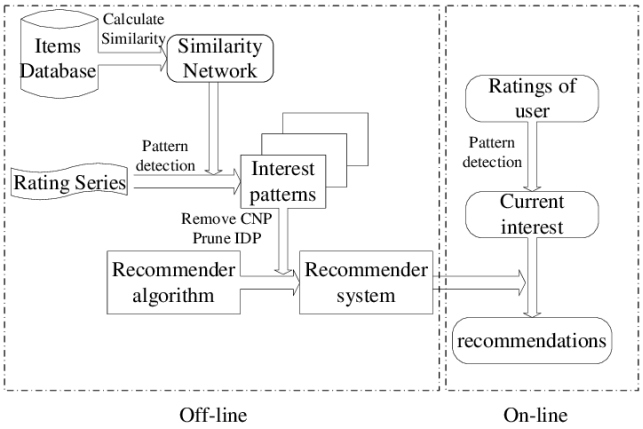

Fig. 4.1: *User interest model framework.*

4. User interest model. Building a user's interest model can give better feedback on the user's interest. A person's hobbies are generally divided into two categories: one is long-term, and the other is short-term. A structural diagram of the user benefit pattern is shown in Figure 4.1.

Long-term interests reflect a person's preference for one thing over a while. This hobby is not something that will change over some time but a constant state. Most of these are gradually accumulated over a long period of life, and naturally, it is also related to the individual's educational experience, life background and personality [15]. Short-term interest usually refers to a specific period. People prefer a thing because of some factors and stimuli, but external stimuli will change this preference. Of course, short-term interests can turn into long-term interests. These fleeting interests will gradually dissipate over time. Because the amount of information users is interested in for a long time is considerable, it is necessary to divide it reasonably. If we use the mathematical formula $S = \{(s_1, \varepsilon_1, r_1), (s_2, \varepsilon_2, r_2), \ldots, (s_n, \varepsilon_n, r_n)\}$ to describe a person's long-term interests, then S_i is what we care about. The ε_i stands for the user's love for the product. r_i indicates that users pay more attention to an event because they prefer it.

Because the short-term interests of users change at any time, it is challenging to transform them into long-term interests, and they are often fleeting, so this paper does not use the method of statistics on the short-term interests of users. The mathematical formula expresses the user's short-term interest in this paper d. β is for something. ξ stands for the level at which users like a product. Then, the user's short-term interest can be expressed by $D = (d_1, d_2, \cdots, d_m)$. In a shopping system based on consumer preferences, this indicator can reflect in real time the change of consumers' preferences for fresh items decreasing over time. We define the expression of freshness as: In a shopping system based on consumer preferences, this indicator can reflect in real time the change of consumers' preferences for fresh items decreasing over time. We define the expression of freshness as:

$$u_t = \begin{cases} 0 & 0 \le u_{t_0} \le \varphi \\ \left(\frac{u_{t_0} - \varphi}{u_{t_0}}\right) & u_{t_0} > \varphi \end{cases}$$

t_0 is the past time. t means now. u_{t_0} indicates that the user has previously found the item novel. u_t represents that the user currently finds the item novel. φ stands for a constant number. If freshness is set to o, then freshness is considered o. Freshness means that it will gradually decrease over time until it reaches o. He has lost interest in the item if the novelty drops to zero. According to different user behavior records, different keywords are automatically generated, and the corresponding database is built for personalized search. However, since the value of the freshness decreases over time, it is also necessary to calculate the weight of the freshness. Then, the calculation of novelty weighting for specific keywords can be expressed by expression.

$$(\beta_p, \xi_1, t_1, h_1, u_1), (\beta_p, \xi_2, t_2, h_2, u_2) \cdots (\beta_p, \xi_n, t_n, h_n, u_n) : h_1, \ldots, h_n$$

The user's recent preferences on a particular tag can be weighted. u is used to represent the total weight. $\xi_p^t = \xi_1 * u_1' + \xi_2 * u_2' \cdots + \xi_n * u_n' \cdot u_1', u_2', \ldots, u_n'$ represents the current freshness value and takes it as the freshness value at t. After weighing each keyword, the interest set can be obtained quickly.

Keywords are weighted according to different user preferences. The detailed calculation method is as follows:

$$S = \{(s_1, \varepsilon_1, r_1), (s_2, \varepsilon_2, r_2), \ldots, (s_n, \varepsilon_n, r_n)\}$$

Since each class has keyword glyph, the weight $s_i = ((\beta_1, w_1), (\beta_2, w_2), \ldots, (\beta_n, w_n))$ of different classes can be obtained by statistical analysis of text feature vectors of different classes. In addition, using the relevant algorithm, the user's interest category can be accurately displayed.

Input: Important bytes that are currently of interest to the user;
Output: the user's long-term interest choice;

$$S = \{(s_1, \varepsilon_1, r_1), (s_2, \varepsilon_2, r_2), \ldots, (s_n, \varepsilon_n, r_n)\}$$

1. Extract all tuples generated on A day t_0 and define them as follows: keyword β, initial weight ξ, creation time t, document containing keyword h, freshness u.
   Firstly, $\xi_p^t = \xi_1 * u_1' + \xi_2 * u_2' \cdots + \xi_n * u_n'$ weighted analysis is carried out for a particular keyword. The corresponding weights of each keyword are obtained after Δt.
2. Repeat step 1 until all tuples have been calculated. Set threshold λ and find a keyword that is larger than this value. Record it in Group ψ. Then, the corresponding text H_0 is determined according to the relationship between the keyword and the tuple.
3. The selection range of the selected associated keywords and text need not be too wide, so according to the threshold value of the initial weight of the keyword, you can find the corresponding collection of files whose original weight exceeds a particular critical value. And as a category according to their level of long-term interest in the category. Use $S_i = (s_1, S_2, \cdots, S_m)\, s_i$ to represent a class.
4. Update keywords when they do not reach a particular range. Re-evaluate it to see if it's of long-term interest.
5. The users' long-term interest set is obtained by calculating steps 3 and 4 . Calculate the number of files r_i in each category to get $((s_1, r_1), (s_2, r_2), \ldots, (s_n, r_n))$. $\varepsilon_i = r_i / \sum r_i$ represents the extent to which users have different preferences for each category. Generate an end-user long-term interest statement:

$$S = \{(s_1, \varepsilon_1, r_1), (s_2, \varepsilon_2, r_2), \ldots, (s_n, \varepsilon_n, r_n)\}$$

5. Design of e-commerce product recommendation system.

5.1. System Architecture. The recommendation of e-commerce products is studied using the calculation method of association rules and user preferences. Figure 5.1 shows the overall architectural design (Picture quoted from Egyptian Informatics Journal, Volume 23, Issue 1, March 2022, Pages 33-45). The system includes data collection and cleaning, user interest analysis, building association rule base, and recommending TOP-N.

(1) Data collection and cleaning: data collection is divided into two parts: shopping and various business activity information. These two data types provide the basis for discovering and researching users' interests. In addition, many messy, fuzzy, incomplete, and dirty data must be cleaned to improve the mining effect. Through the detection of empty orders, goods in the transaction record are detected, and non-relevant fields are removed. In this way, the data is simplified.

(2) User interest analysis module: This module is mainly used to model and update the interest model. It is automatically collected by the front end and stored in the corresponding database. A "user-interest item" chain is constructed using the above calculation method. Search for items with high correlation and apply them to the TOP-N model.

(3) Establishment of association rule database model: the CTE-MARM method proposed above is used to discover the connections between specific levels in specific application scenarios. Set the restriction level k to 2. The rules stored in the transaction table are classified from different perspectives according to credibility.

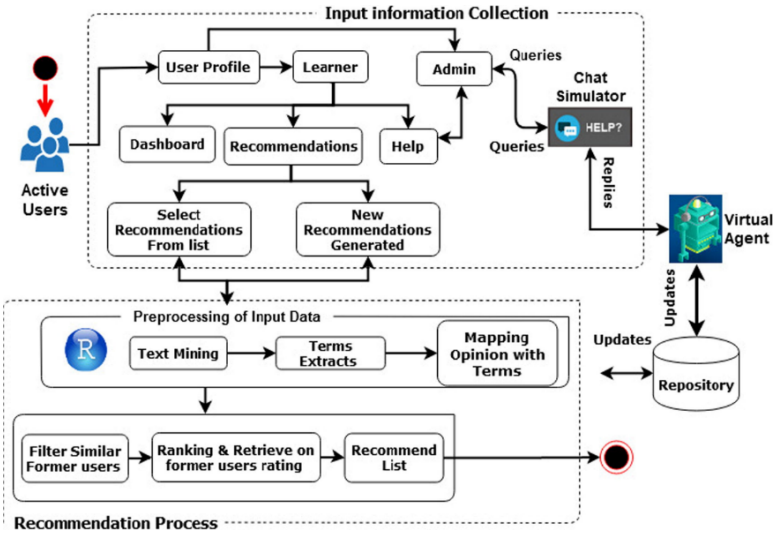

Fig. 5.1: *Overall architecture diagram of the system.*

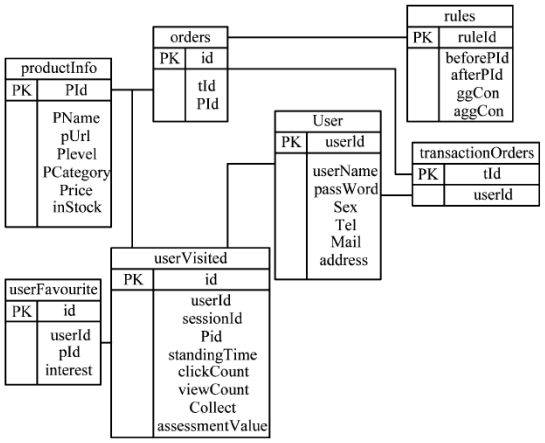

Fig. 5.2: *Database E-R diagram.*

(4) TOP-N network recommendation mode: The page is visually displayed through the association rules of the products that correlate more with the user's interest mode.

5.2. Database Design. According to the characteristics of e-commerce product recommendation, a database E-R model based on user behavior and product recommendation is proposed. The database diagram for E-R is shown in Figure 5.2.

Below are the main database tables.

1. The user information form includes user ID, registration name, password, gender, contact information, address, etc.
2. The item information form includes item ID, name, link, grade, classification, price, inventory, etc.
3. The transaction table contains the transaction ID and user ID.
4. The purchase record form contains the unique identification ID, transaction processing ID and item ID.
5. The user behavior table contains unique identifiers, user identifiers, item identifiers, number of clicks,

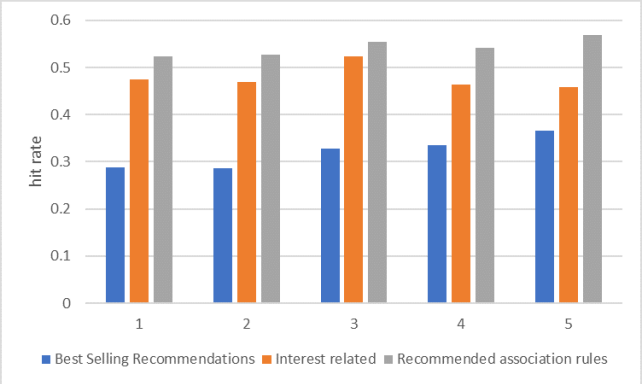

Fig. 6.1: *Hit ratio analysis of manual analysis and system algorithm results.*

repeat visits, browsing time, and whether to collect and evaluate scores.

6. The association rule table contains a unique identification, pre-rule component identification, post-rule component identification, inter-item confidence, classification and inter-item confidence.
7. The user interest table contains unique identifiers, user identifiers, product identifiers and interest levels.

6. System testing and verification. The shopping data of 3000 users are used as experimental data to test the accuracy of the established theory and products. The results of manual statistics are compared with user habits and system operation results of manual analysis. Cross-merge was performed after each trial. There were five trials. The result is shown in Figure 6.1. The study found that the accuracy of manual analysis of the best-selling items reached 31.8%. The interest recommendation accuracy of manual analysis reached 51.1%, while the recommendation accuracy of the e-commerce product recommendation system based on association rules reached 55.8%, which is more efficient and accurate than the method. This achieves the original design goal.

7. Conclusion. This paper uses the method based on CTE-MARM to establish an e-commerce data analysis method based on user interests and association rules. In this way, TOP-N products are recommended for users. However, the multi-level association discovery method research needs to be further explored. More influencing factors and technical means must be considered in improving the hit rate of product recommendations.

8. Acknowledgments. The work was supported by 1. Industry and University Collaborative Education Project of the Ministry of Education in 2024: Construction of University about E-commerce Big-Data Practice Base From the perspective of intelligent new media (No. 231100273283538). 2. Industry and University Collaborative Education Project of the Ministry of Education in 2022: Construction of University about E-commerce Big-Data Practice Base from the Perspective of New Media (No. 220606030204538).

REFERENCES

[1] Dogan, O., Kem, F. C., & Oztaysi, B. (2022). Fuzzy association rule mining approach to identify e-commerce product association considering sales amount. Complex & Intelligent Systems, 8(2), 1551-1560.
[2] Yang, W., & Lin, Y. (2021). Research on the interactive operations research model of e-commerce tourism resources business based on big data and circular economy concept. Journal of Enterprise Information Management, 35(4/5), 1348-1373.
[3] Zong, K., Yuan, Y., Montenegro-Marin, C. E., & Kadry, S. N. (2021). Or-based intelligent decision support system for e-commerce. Journal of Theoretical and Applied Electronic Commerce Research, 16(4), 1150-1164.
[4] Chawla, N., & Kumar, B. (2022). E-commerce and consumer protection in India: the emerging trend. Journal of Business Ethics, 180(2), 581-604.
[5] Abendin, S., & Duan, P. (2021). Global E-commerce talks at the WTO: Positions on selected issues of the United States, European Union, China, and Japan. World Trade Review, 20(5), 707-724.
[6] Ayob, A. H. (2021). E-commerce adoption in ASEAN: who and where? Future business journal, 7(1), 1-11.

[7] Hasana, Z., & Afifah, I. I. (2023). Perlindungan hukum terhadap konsumen dalam transaksi e-commerce. Advanced In Social Humanities Research, 1(5), 795-807.
[8] Karn, A. L., Karna, R. K., Kondamudi, B. R., Bagale, G., Pustokhin, D. A., Pustokhina, I. V., & Sengan, S. (2023). Customer centric hybrid recommendation system for E-Commerce applications by integrating hybrid sentiment analysis. Electronic Commerce Research, 23(1), 279-314.
[9] Bawack, R. E., Wamba, S. F., Carillo, K. D. A., & Akter, S. (2022). Artificial intelligence in E-Commerce: a bibliometric study and literature review. Electronic markets, 32(1), 297-338.
[10] Aditantri, R., Mahliza, F., & Wibisono, A. D. (2021). Urban planning and e-commerce: Understanding the impact during pandemic covid-19 in Jakarta. International Journal of Business, Economics, and Social Development, 2(3), 135-142.
[11] Belwal, R., Al Shibli, R., & Belwal, S. (2021). Consumer protection and electronic commerce in the Sultanate of Oman. Journal of Information, Communication and Ethics in Society, 19(1), 38-60.
[12] Peráček, T. (2022). E-commerce and its limits in the context of the consumer protection: The case of the Slovak Republic. Tribuna Juridică, 12(1), 35-50.
[13] Tran, D. T., & Huh, J. H. (2022). Building a model to exploit association rules and analyze purchasing behavior based on rough set theory. The Journal of Supercomputing, 78(8), 11051-11091.
[14] Ünvan, Y. A. (2021). Market basket analysis with association rules. Communications in Statistics-Theory and Methods, 50(7), 1615-1628.
[15] Cui, H., Niu, S., Li, K., Shi, C., Shao, S., & Gao, Z. (2021). A k-means++ based user classification method for social e-commerce. Intell. Autom. Soft Comput, 28(1), 277-291.

Edited by: Zhigao Zheng
Special issue on: Graph Powered Big Aerospace Data Processing
Received: Dec 16, 2023
Accepted: Jan 9, 2024