Hindawi
Complexity
Volume 2019, Article ID 6543590, 8 pages
https://doi.org/10.1155/2019/6543590

Research Article

EC-Structure: Establishing Consumption Structure through Mining E-Commerce Data to Discover Consumption Upgrade

Lin Guo[1] **and Dongliang Zhang**[2]

[1]*School of Economics and Management, Changchun University of Science and Technology, Jilin 130022, China*
[2]*Institution of technical science, Fudan University, Shanghai 200000, China*

Correspondence should be addressed to Lin Guo; guolin@cust.edu.cn

Received 24 December 2018; Accepted 26 February 2019; Published 12 March 2019

Guest Editor: Thiago C. Silva

The traditional methods of analyzing consumption structure have many limitations, and data acquisition is difficult, so it is hard to scientifically verify the accuracy of algorithms. With the development of Internet economy, many scientific researchers focus on mining knowledge of consumer behavior using big data analysis technology. Because consumption decisions are influenced by not only personal characteristics but also social trends and environment, it is one-sided to analyze the impact of one single factor on the phenomenon of consumption. The authors of this paper combine the consumption structure analysis method and data processing technology using data from an e-commerce platform to extract the consumption structure of cities, compare the structural differences between different periods, and then discover consumption upgrading according to swarm intelligence. The experiments prove the efficacy of the algorithm proposed in this paper compared to other similar algorithms using several different datasets, which illustrates the algorithm's efficacy and stable performance in consumption structure analysis.

1. Introduction

With the continuous expansion of consumption scale, consumers' personalized demands are becoming increasingly obvious. Consumer behaviors, such as purchasing decisions, are influenced by not only personal characteristics but also interpersonal relationships, social environments, network culture, and so on. Therefore, the analysis of consumption just from the individual perspective is one-sided and unscientific.

At present, most research on consumption upgrading is carried out from the macroperspective and is realized based on national statistical yearbook data. The analysis of consumption upgrading from a macroperspective is relatively simple. However, it is difficult to find individual consumption structures at the microlevel from the perspective of consumers. From the perspective of management and economics, consumption upgrading is difficult to measure, and there is no strict boundary with which to distinguish between consumption upgrading and nonupgrading, and relevant experimental data is also difficult to obtain. The authors of this paper can fully quantify the judgment process of consumption upgrading, propose a set of evaluation criteria, and use big data with comprehensive coverage of user features for mining. Therefore, the algorithm proposed in this paper is scientific and accurate.

This paper combines the consumption structure analysis method and data processing technology to extract collective wisdom to construct an economic map that describes urban economic hotspots. The algorithm involves studying consumer consumption structures to realize consumption upgrading mechanism research and building a consumption upgrading model to analyze whether consumption upgrading occurred. The results obtained from the multiangle and multidimensional research will be comprehensive and reasonable.

2. Related Work

The research on consumption function theory is mainly focused on Persistent Income Theory (PIH) and Life Cycle Theory (LCH). The only difference between PIH and LCH is

that the former usually uses an indefinite limitation while the latter uses a definite limitation, so they are generally called LC-PIH when combined. Many researchers [1–4] use LC-PIH to study the consumption problems of Western residents, but their conclusions are inconsistent.

With the development of the Internet economy, an increasing number of researchers focus on the impact of Internet technology. Electronic commerce (e-commerce) is any type of business or commercial transaction that involves information transfer across the Internet [5]. A user's behaviors on a website can reflect their interests and purchase intentions. Therefore, consumption structure can be analyzed by studying the data of e-commerce platforms, which is difficult to realize using to traditional research methods.

When analyzing the characteristics of a network, the distribution of user activity is investigated and a network of bidders that is connected by common interest in individual articles is constructed [6, 7]. The network's cluster structure corresponds with the main user groups according to common interests, exhibiting hierarchy and overlap.

Regarding the characteristics of uses, Curme [5], Chattopadhyay [8], and Guo [9] analyze user behaviors from the perspective of complex systems and extract implicit semantic information from large-scale semistructured data. Glass [10], Singh [11], and Aviano [12] find personal characteristics through the feature extraction of original website data, mine important features of users using to feature selection methods, and finally extract a user profile model. Kim [13], Ouaftouh [14], and Diao [15] classify customers into different groups according to their similarities that are calculated through demographic features and psychological features or features such as customer value, customer consumption, and lifetime value.

3. Classification of Consumption Data

The rapid development of social networking resulted in the explosive growth of data that contains large amounts of high-quality information, such as information about user interests and interpersonal relationships. Therefore, social data processing and knowledge mining are intricate and indispensable. Consumption data contains much redundant or nonrepresentative data. We adopt the matrix analysis method to quickly classify consumption data into negative data, positive data, insufficient evidence data, and disputed data.

Definition 1 (negative data). The normalized value of negative comment times is greater than 0, and the normalized value of positive comment times is less than 0.

Definition 2 (positive data). The normalized value of positive comment times is greater than 0, and the normalized value of negative comment times is less than 0.

Definition 3 (insufficient data). The number of positive and negative comments is relatively small, so there is no authoritative data based on which to measure the nature of the data.

Definition 4 (controversial data). The number of positive and negative comments is relatively large, so there is no way to find absolute, salient features based on which to measure the nature of the data.

Because socialized data can reflect the characteristics and interpersonal relationship information about real society, the knowledge of regional and overall consumption structure (the coverage of analysis results are determined by data capture granularity) can be acquired by analyzing the data of <user, comment> that is gathered from websites. After word segmentation and semantic analysis, we can obtain the <consumption object, comment times, positive times, negative times> data. Due to large differences in the evaluation data about different consumption objects, we use standardized data to control data to within a range to measure the popularity of different consumption objects. The formula for data normalization is shown as follows:

$$(x, y) = \left(\frac{x_i - \overline{x}}{s_x}, \frac{y_i - \overline{y}}{s_y} \right) \tag{1}$$

$$\overline{x} = \frac{1}{n} \sum_{p=1}^{n} x_p \tag{2}$$

$$\overline{y} = \frac{1}{n} \sum_{p=1}^{n} y_p \tag{3}$$

$$s_x = \sqrt{\frac{1}{n-1} \sum_{p=1}^{n} \left(x_p - \overline{x} \right)^2} \tag{4}$$

$$s_y = \sqrt{\frac{1}{n-1} \sum_{p=1}^{n} \left(y_p - \overline{y} \right)^2} \tag{5}$$

(x, y) is the calculated result after standardization. Its mean value is 0, the variance is 1, and it is dimensionless. (x, y) can be mapped to a two-dimensional coordinate interval [-1, +1]. The standardized variable value fluctuates around 0. A value greater than 0 indicates that (x, y) is higher than the average level, and a value less than 0 indicates that (x, y) is lower than the average level.

The authors of this paper use nodes to describe consumption objects. Therefore, the locations of nodes in the matrix can describe status of consumption objects. In the consumption matrix, X and Y coordinates, respectively, represent the standardized data of positive data and negative data. In the four interval matrixes, the consumption data is divided into four categories.

The consumption matrix is divided into four regions, as shown in Figure 1.

It can be seen from Figure 1 that the X-coordinate value of the node in the negative partition is less than 0 and the Y-coordinate value is greater than 0, which indicates that the nodes in the negative partition have more negative data than the average. The nodes in the positive partition are just the opposite. The other two types of nodes are controversial and insufficient nodes. Among them, the controversial node

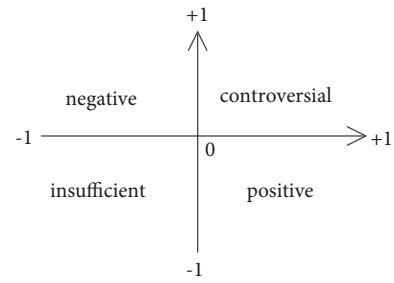


Figure 1: Coordinate distribution of four types of nodes.

has much positive and negative information. The insufficient node has little positive and negative information. Neither of these two types of nodes can be classified into one specific partition. To reduce the impact of invalid or redundant data on the accuracy of the analysis, the authors of this paper only focus on the nodes in positive and negative areas. In addition, by manually analyzing the nodes in positive and negative regions, it is found that the data about these nodes is authoritative and clean. It is enough to describe the information about consumption structure of users and sufficient for the subsequent experiments.

4. Analysis of Consumption Coefficient

By analyzing the data obtained in the above process, the matrix ITEM(x, y) is constructed from positive and negative data. Each node in the matrix describes the situation of the positive and negative comments about a consumption object. By comparing the consumption matrixes from different periods, the consumption trends, changes to the structure, and consumption upgrading can be judged and identified.

With the continuous expansion of consumption scale, consumers' personalized demands are becoming increasingly obvious. Consumer behavior, such as purchase decisions, is affected by multiple factors, so consumption hotspots and structures often change. This change may be strong or weak; of course significant changes in structures are relatively easy to detect, but weak changes are difficult to capture. Therefore, to identify changes in consumption structures, it is necessary to calculate the proportions of different consumption objects in the total consumption field and discover changes of consumption structure in time by analyzing changes in proportion. In this paper, we calculate the consumption coefficient of each consumption object, measure the proportion of different consumption objects in the total consumption field, and then identify changes and trends in the consumption structures of users.

The formula for calculating the consumption coefficient is as follows:

$$coeff = \frac{heat\,(q)}{heat\,(all)} \times 100\% \tag{6}$$

coeff is the consumption coefficient of consumption object q. heat(q) represents the heat of the consumption object q. heat(*all*) represents the heat of all consumption objects. It can be seen from the formula that the consumption coefficient is the proportion of a certain consumption object in the total consumption objects. The consumption coefficient is added to the matrix ITEM(x, y) as an additional parameter, so the expression of the matrix becomes ITEM(x, y, *coeff*). Therefore, the structural characteristics of consumption can be described from three dimensions. By analyzing ITEM(x, y, *coeff*), the implied information about consumption structure, consumption trend, and consumption upgrading can be obtained. The detailed analysis process is described in the next section.

5. Discovery of Individual Consumption Upgrading

To compare the consumption data from different periods, the differing degrees of the matrixes that describe the consumption structures in different periods need to be calculated. If the difference degree exceeds a certain threshold value, then the consumption upgrading phenomenon is considered as happening. Here, the consumption matrix at moment n is denoted as ITEM_n and the consumption matrix at moment n+1 is denoted as ITEM_{n+1}. By comparing ITEM_n and ITEM_{n+1}, the differences of different consumption matrixes can be calculated and structural changes can be detected. The formula for calculating degree of difference is as follows:

$$COR\left(\mathrm{ITEM}_n, \mathrm{ITEM}_{n+1}\right) = \frac{\sum_{i=1}^{n}\left(ITEM_i - \overline{ITEM_n}\right)\sum_{j=1}^{n+1}\left(ITEM_j - \overline{ITEM_{n+1}}\right)}{\sqrt{\sum_{i=1}^{n}\left(ITEM_i - \overline{ITEM_n}\right)^2 \sum_{j=1}^{n+1}\left(ITEM_j - \overline{ITEM_{n+1}}\right)^2}} \tag{7}$$

According to the formula, the coefficient *COR* is obtained by dividing the covariance by the standard deviation of two variables. The covariance can reflect the correlation degree between two random variables. When the covariance is greater than 0, it means that the two variables are positively correlated, and when the covariance is less than 0, it means that the two variables are negatively correlated. Note that the coefficient is meaningful when both variables are not zero, and the range of the coefficient is [-1, 1]. When *COR* is 1, ITEM_n and ITEM_{n+1} are completely positively correlated. When *COR* is -1, ITEM_n and ITEM_{n+1} are completely

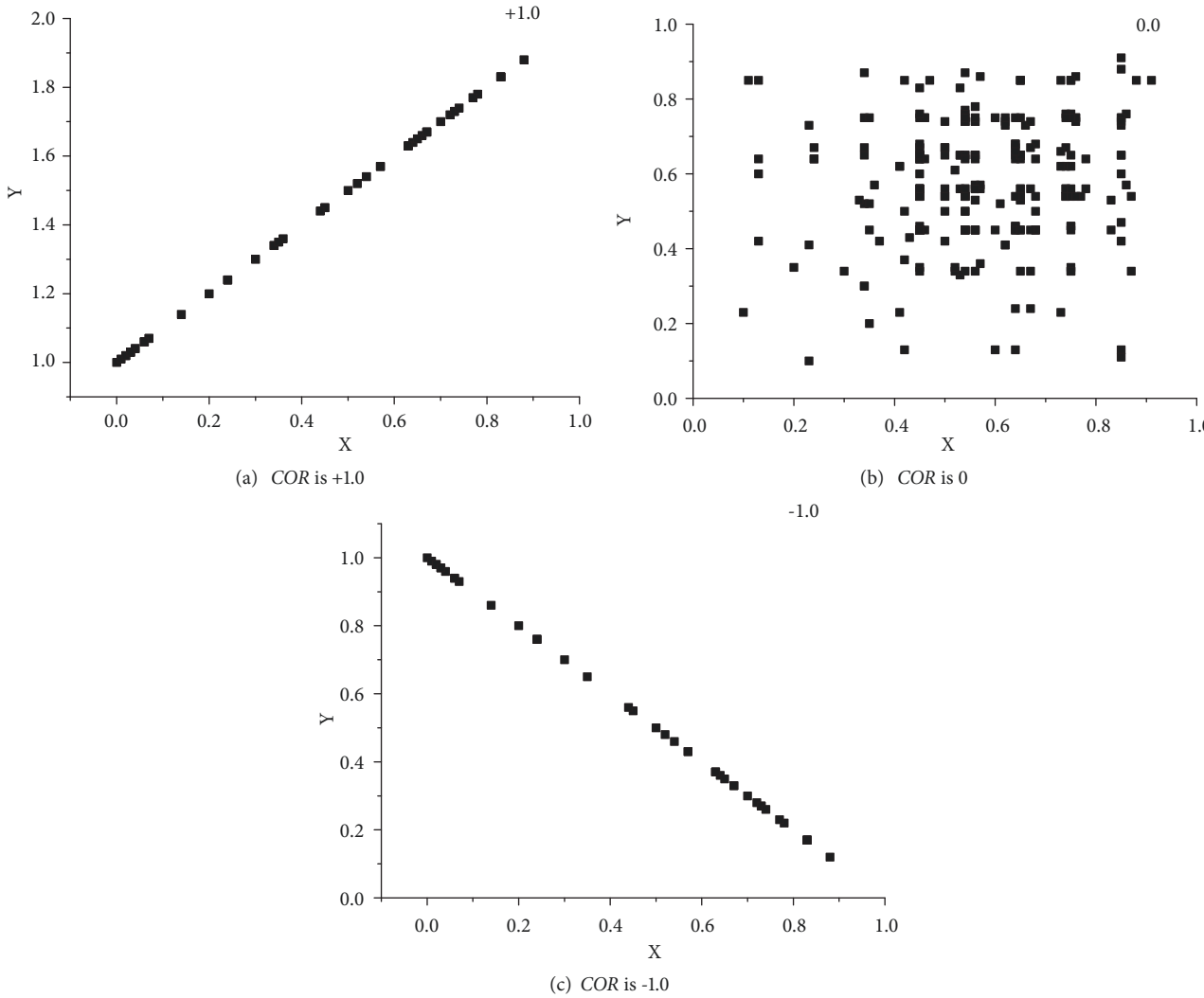


Figure 2: The coefficient *COR* of different consumption structure.

negatively correlated. The greater the absolute value of *COR* is, the stronger the correlation degree between $ITEM_n$ and $ITEM_{n+1}$ is. The closer the coefficient *COR* is to 0, the weaker the correlation degree between $ITEM_n$ and $ITEM_{n+1}$ is.

Through the above methods, we can build consumption matrixes for different periods and judge the differences in consumption structure in between periods by calculating the coefficient, *COR*. When *COR* approaches 1 or -1, it means that the consumption structure significantly changed, so it can be considered that consumption upgrading occurred. As shown in Figure 2, the closer the coefficient is 1 or -1, the greater the structural difference is, while a coefficient that is near 0 indicates that the consumption structure changed little and there is no upgrading.

6. Experiment

The datasets used throughout the experiments are Zachary's Karate Club(http://www-personal.umich.edu/~mejn/netdata/), Dolphin's Associations(http://www-personal.umich.edu/~mejn/netdata/), LesMiserables(http://wiki.gephi.org/index.php/Datasets), MovieLens(http://www.datatang.com/datares/detail.aspx?id=44295), and EP dataset(http://www.dianping.com/).

(1) The dataset of Zachary's Karate Club is a social network of friendships between 34 members, so edges in the graph describe the higher frequency of interactions between members.

(2) The dataset of Dolphin's Associations is an undirected social network of frequent associations between 62 dolphins, which has 62 nodes and 159 edges.

(3) The dataset of LesMiserables is a coappearance network of characters in Les Miserables, which contains 77 nodes and 254 edges.

(4) The dataset of MovieLens is a synthesized recommendation system and virtual community, which is commonly used for social computing.

(5) The EP dataset was captured from an e-commerce platform (dianping.com). It contains 15,890,209 pieces of data and was updated in August 2018. The data collection fields are

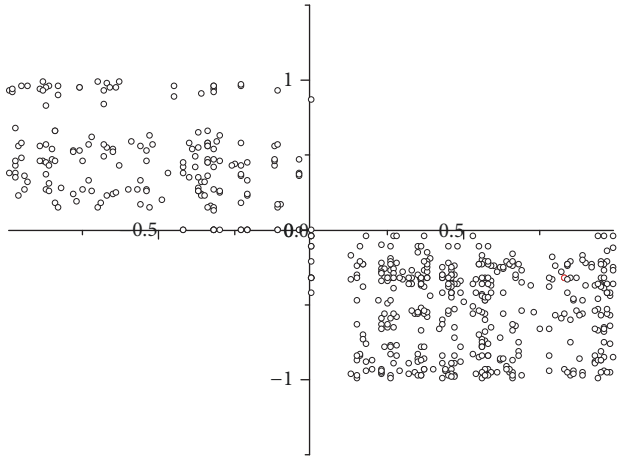


Figure 3: The positive and negative distribution of comments.

shop_id (uniqueness), province, city, city_id, area, big_cate (the primary classification), big_cate_id, small_cate (the secondary classification), small_cate_id, service_rating, all_remarks, very_good_remarks (5-star review), good_remarks (4-star review), common_remarks (3-star review), bad_remarks (2-star review), and very_bad_remarks (1-star review).

Comparison Methods. NMFOSC [16] presents an approach to community detection that utilizes a nonnegative matrix factorization model to divide overlapping communities from networks. RNM [17] is a local expansion method based on rough neighborhood. CPM [18] greedily expands natural communities of seeds until the whole graph is covered by using a local fitness function. EdgeB-Cluster [19] bundles similar edges, adjusts the locations of nodes to optimize the visualized output of the graph and analyzes networks from a community level.

Through the analysis of a consumption object in a certain region in the e-commerce platform, it was found that the number of positive comments is very large. This is because there is a phenomenon of deliberately increasing the number of good comments to improve the store's reputation, which results in the presence of too many good comments. On the contrary, the numbers of neutral and negative comments are relatively reasonable, and few of these comments are intentionally added or deleted, so they are convincing. Based on the above factors, the authors of this paper did not analyze the quantity of positive comments and only considered the quantity of neutral and negative comments. Through experimental verification of the quality of the neutral and negative comments, it was found that the data is authentic and abundant, and enough to describe the object to be tested.

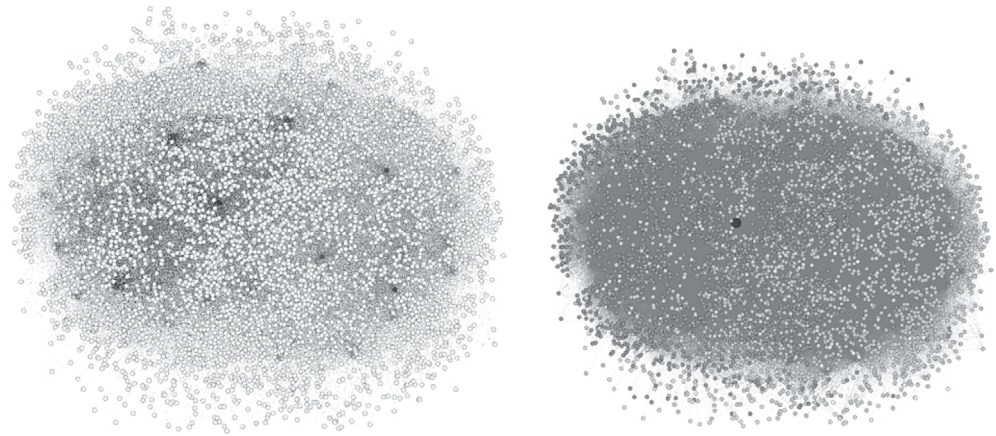
Figure 3 shows the distribution of positive and negative comments. The nodes in the insufficient and controversial areas do not provide valuable information for subsequent analysis, so they were removed. It can be seen that there are more positive nodes than negative ones, and the difference between them is large. It is important to note that although the nodes in the negative area represent that the user's comments are negative, they still provide useful knowledge about consumption trends that cannot be removed.

Figure 3 shows the regional analysis results while Figure 4 shows the overall analysis results. Figure 4 is the visual output of the results generated by the algorithm. Figure 4(a) shows the distribution of nodes, and the colors of nodes indicate the heat of different consumption objects. The darker a node color is, the more attention the node received. Figure 4(b) is a heat map of a center node, and the black node is the center node. Figure 4(b) shows that there is a certain correlation between the central node and a large number of other nodes, indicating that there are many high correlations between different consumer groups. Thus, the characteristics of consumption objects can be further analyzed based on the relationships between consumers and commodities.

Figure 5 depicts the distribution of different consumption objects. In this case, the node with a ratio of more than 0.6 is regarded as a popular consumption node while a node with a ratio of less than or equal to 0.6 is regarded as an unpopular consumption node. Of course, if the ratio threshold is lowered, then additional nodes will be divided into popular consumption areas. It can be seen from Figure 5 that most nodes belong to the nonhot field, which is in-line with the actual situation.

Figure 6(a) depicts the characteristics of nodes that were divided into two categories to describe different consumption heat (some representative nodes are extracted). It can be seen that the characteristics of nodes in different categories vary greatly. Figure 6(b) describes the closeness centrality distribution of the nodes belong to the same category. This shows that the node locations have normal distribution, so the similarity between the nodes in the same category is very high. That is to say, the classification is reasonable.

For the purpose of analyzing the experimental results, the following measurement parameters are used [20]: Multiplicity Precision calculated by $MP = \min(|C(e) \cap C(e')|, |L(e) \cap L(e')|)/|C(e) \cap C(e')|$, Multiplicity Recall by $MR = \min(|C(e) \cap C(e')|, |L(e) \cap L(e')|)/|L(e) \cap L(e')|$. Let $L(e)$ and



(a) Distribution of nodes (b) The heat map

Figure 4: Visualization of the results generated by the algorithm.

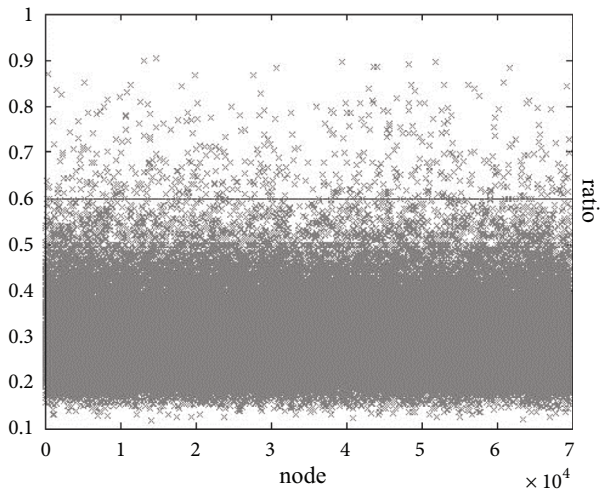


Figure 5: The distribution of different consumption objects.

$C(e)$ denote the category and the cluster of an item e. e is a cluster with n items belonging to the same category and e' is a cluster merging n items from unary categories. FB is a comprehensive measure of MP and MR, and the algorithm is $FB=MP\times MR\times 2/(MP+MR)$.

Table 1 proves the validity and feasibility of the algorithm. The numbers in italic indicate the highest value of the same parameter in each row. Table 1 displays a comparison of the algorithm proposed in this paper to other similar algorithms. The datasets Karate Club, Dolphin, LesMiserables, and MovieLens are used to prove the performance of the algorithms in structural analysis. The EP dataset is used to prove the performance of e-commerce data analysis. It is found that EC-Structure performs better than other algorithms and performs stably with different data sets. The main reasons for which EC-Structure is superior to other algorithms are that (1) it reduces the influence of erroneous e-business platform data on the algorithm, (2) it increases the consumption coefficient as a parameter with which to measure the proportions of different consumption objects, and (3) the coefficient COR can help researchers accurately judge changes in consumption structures. Therefore, the operation effect of this algorithm is effective.

7. Conclusion

Research on consumer behavior can be made by extracting and analyzing useful information from a large amount of incomplete, vague, and random consumer behavior data. The algorithm proposed in this paper builds consumption structures and a consumption upgrading model based on the data from e-commerce platforms to analyze whether consumption upgrading occurred. The results of the experiment verified the implementation efficacy and analysis accuracy of the algorithm. It was found that the algorithm is effective. The implementation efficacy of the proposed algorithm is superior to those of other algorithms, and it runs stably with different datasets.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

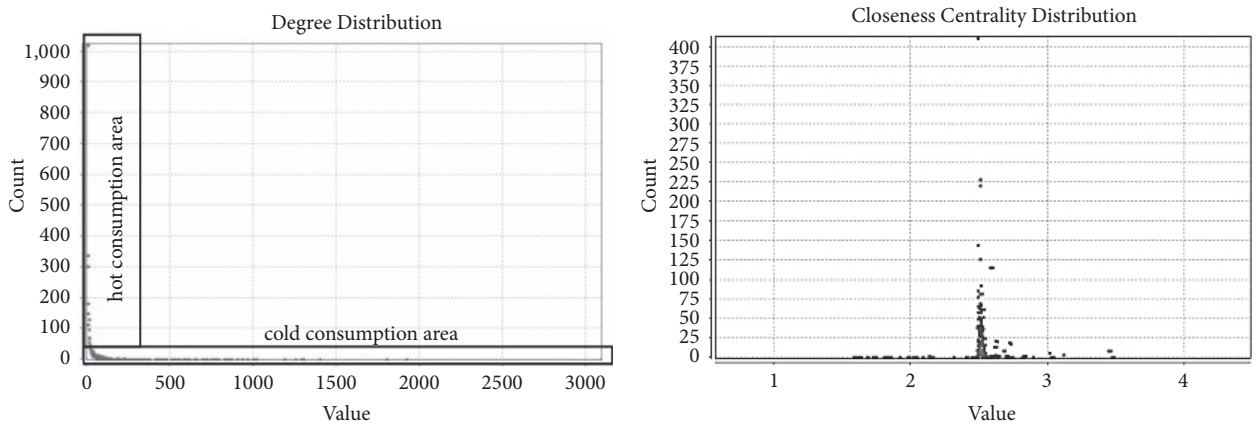
The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by Youth Fund of Humanity and Social Science of Ministry of Education of China (Grant no. 18YJCZH041); Project of Education Department of Jilin Province of China (Grant no. JJKH20190612SK).

Table 1: The performance comparisons.

dataset	NMFOSC			RNM			CPM			EdgeB-Cluster			EC-Structure		
	MR	MP	FB	MR	MP	FB	MR	MP	FB	MR	MP	FB	MR	MP	FB
Karate Club	*1.00*	0.92	0.96	0.84	*1.00*	0.91	0.58	0.94	0.71	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*
Dolphin	0.64	0.90	0.75	0.46	0.97	0.62	0.40	0.94	0.56	0.73	*0.98*	0.83	*0.80*	*0.98*	*0.88*
LesMiserables	0.72	0.87	0.79	0.80	0.88	*0.84*	0.48	*0.89*	0.62	*0.81*	0.88	*0.84*	*0.88*	0.83	*0.85*
MovieLens	*0.83*	0.85	0.84	0.56	0.86	0.68	0.81	0.86	0.83	0.81	*0.88*	*0.84*	0.82	*0.88*	*0.85*
EP dataset	*0.85*	0.80	0.82	0.53	0.56	0.54	0.53	0.65	0.58	0.79	*0.83*	0.81	0.85	0.82	*0.83*



(a) Degree distribution (b) Closeness centrality distribution

Figure 6: The characteristic distribution of hot consumption and cold consumption nodes.

References

[1] P. N. Ireland, "Using the permanent income hypothesis for forecasting," *Federal Reserve Bank of Richmond Economic Quarterly*, vol. 81, no. 1, pp. 49–63, 1995.

[2] L. A. Fisher and G. Kingston, "Improved forecasts of tax revenue via the permanent income hypothesis," *Australian Economic Review*, vol. 50, no. 1, pp. 21–31, 2017.

[3] L. Zhou, C. Wang, and S. O. Finance, "Household debt and consumption-evidence from micro data," *Soft Science*, vol. 3, pp. 32–43, 2018.

[4] M. Zagler, "Empirical evidence on growth and business cycles," *Empirica*, vol. 44, pp. 1–20, 2017.

[5] C. Curme, T. Preis, and H. E. Stanley, "Quantifying the semantics of search behavior before stock market moves," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 111, no. 32, pp. 11600–11605, 2014.

[6] J. Reichardt and S. Bornholdt, Ebay users from stable groups of common interest, 2005.

[7] H. Halpin, "The semantics of search," in *Social Semantics*, pp. 149–186, Springer US, 2013.

[8] T. Chattopadhyay, S. Maiti, A. Pal et al., "Automatic discovery of emerging trends using cluster name synthesis on user consumption data: extended abstract," in *Proceedings of International Conference Companion on World Wide Web*, pp. 981–983, 2016.

[9] L. Guo, W. Zuo, and T. Peng, "Inference network building and movements prediction based on analysis of induced dependencies," *IET Software*, vol. 11, no. 1, pp. 12–17, 2017.

[10] B. Glass, Z. Benenson, and R. Landwirth, "Look before you leap: improving the users' ability to detect fraud in electronic marketplaces," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 3870–3882, ACM, 2016.

[11] P. Singh and M. Singh, "Fraud detection by monitoring customer behavior and activities," *Annals of Regional Science*, vol. 49, no. 1, pp. 1–27, 2012.

[12] D. Aviano, B. L. Putro, and E. P. Nugroho, "Behavioral tracking analysis on learning management system with apriori association rules algorithm," in *Proceedings of the 2017 3rd International Conference on Science in Information Technology (ICSITech)*, Bandung, Indonesia, 2017.

[13] K. Kim, Y. Choi, and J. Park, "Pricing fraud detection in online shopping malls using a finite mixture model," *Electronic Commerce Research and Applications*, vol. 12, no. 3, pp. 195–207, 2013.

[14] S. Ouaftouh, A. Zellou, and A. Idri, "User profile model: a user dimension based classification," in *Proceedings of the 2015 10th International Conference on Intelligent Systems: Theories and Applications (SITA)*, Rabat, Morocco, 2015.

[15] Y. Diao, K. Y. Liu, and L. Hu, "Classification of massive user load characteristics in distribution network based on agglomerative hierarchical algorithm," in *Proceedings of the 2016 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*, Chengdu, China, 2017.

[16] N. Chen, Y. Liu, and H.-C. Chao, "Overlapping community detection using non-negative matrix factorization with orthogonal and sparseness constraints," *IEEE Access*, vol. 6, pp. 21266–21274, 2017.

[17] Z. H. Zhang, D. Q. Miao, and J. Qian, "Detecting overlapping communities with heuristic expansion method based on rough neighborhood," *Chinese Journal of Computer*, vol. 36, no. 10, 2013.

[18] F. Havemann, M. Heinz, and A. Struch, "Identication of overlapping communities and their hierarchy by locally calculating community-changing resolution levels," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 1, 2011.

[19] L. Guo, W. Zuo, T. Peng, and B. K. Adhikari, "Attribute-based edge bundling for visualizing social networks," *Physica A: Statistical Mechanics and Its Applications*, vol. 438, pp. 48–55, 2015.

[20] E. Amigó, J. Gonzalo, J. Artiles, and F. Verdejo, "A comparison of extrinsic clustering evaluation metrics based on formal constraints," *Information Retrieval*, vol. 12, no. 4, pp. 461–486, 2009.