2022 IEEE 5th International Conference on Computer and Communication Engineering Technology (CCET) | 978-1-6654-6735-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/CCET55412.2022.9906375

Research on a User Context Model Based on Data Mining

Jinhai Li, Lihang Ling, Yue Wu
Taizhou University
Taizhou 225300, China
e-mail: ljh-hk@163.com

***Abstract*—With the popularity of the Internet and under the rising number of online shoppers, e-commerce platforms are facing increasing challenges in personalizing their descriptions and information services to users. Therefore, in this paper, the context factors that have an impact on users' decisions are counted through questionnaires and derived by way of statistical analysis. Based on the coarse-grained context conclusions drawn from the statistical analysis through SPSS, a web crawler was used to crawl the Taobao users' context and behavioral dataset as the experimental dataset. The crawled data was then analyzed by K-Means. Finally, the K-Means clustering model analysis is used to propose the construction of a user context model, which is used to promote the correct understanding of personalized information needs on the platform, thereby increasing user loyalty.**

Keywords-user context; K-Means; personalized service; SPSS; web crawler

I. INTRODUCTION

With the wide use of mobile intelligent devices, users are faced with more and more information, and a large amount of data generated in users' daily life is also recorded by software. Faced with much information, personalized recommendation technology can effectively solve these problems. It automatically collects the user's information, such as gender, age, click times, stay time and other behaviors. The relevant modeling technology is used to analyze users' behavior and information, process the behavior rules and information of users, establish a user model, and predict the possibility of current users to purchase products. Finally, it displays this information to users through filtering.

User modeling is the foundation and core of personalized recommendation [1]. A well-considered user model can greatly improve the quality of the recommendation system, better understand the personalized needs of users, and improve the user experience and quality, thus stimulating the consumption behavior of users to increase the income of enterprises [2]. For example, 35 percent of Amazon's product sales are provided by its Amazon Recommendation System [3]. To build a good user model is not only to integrate the mined data, but also to integrate the context. For example, the season, weather, mood and so on in the user's location will affect his choice of things, which is the difficulty in the construction of a user context model. However, most of the current user models are too monotonous and lack of timeliness. They can only record and model through the user's past browsing data and the basic information of user registration. The commodities predicted by such user models are too limited, lacking timeliness and unable to meet the daily personalized needs of users. This is the opportunity and challenge of the current user context model.

II. RELATED RESEARCH REVIEW

The earliest study on the concept of context began in 1950. HermanKahn [4] introduced concepts such as context analysis into the fields of planning and strategy. Scientists and scholars are more interested in the research of network technology and information science. Associine for Computer Mair coolinghinery held four consecutive conferences on Situational Attention Metadata from 2006 to 2009.With the in-depth and comprehensive study of the context, Hong et al. divided the context awareness system into the basic research layer, the network layer, the middleware layer, the application layer and the user infrastructure layer. In addition, they proposed the framework of providing personalized information services based on the context awareness computing by using the situation history and built a prototype system [5]. Gavalas et al. made recommendations based on collaborative filtering algorithm by using information of similar tourists and combining situational information (location, time, weather conditions, visited scenic spots, etc.) [6].

In practice, applications are more extensive, such as e-commerce, advertising, film recommendation, music recommendation, library services, mobile learning and other fields all have had in-depth development. Amomavicius et al. proposed a tourism recommendation system, which combined pre-context filtering, post-context filtering and modeling to make context awareness recommendation for users, and provided different vacation choices based on different locations and seasons [7]. In the field of library, Minrva App developed by Hahn can find relevant resources near the current user's location, and then recommend relevant resources and provide corresponding locations to the user [8].

In the field of computer, Yue et al. proposed a context-aware and scheduling strategy, and established an intelligent interactive system structural model based on context-aware [9]. Chen and Liu proposed an activity-based context awareness model, on this basis, they proposed an activity-based interaction design method of context-aware system [10].

Research on the user context model. Ge et al. used M-C-W user interest to model and used the inflow of context data and network data to capture, identify, classify and process information. Finally, they found that the accuracy of M-C-W

user interest model algorithm was higher than the algorithm based on user interest, project and push [11].

To sum up, the current research mainly completes personalized recommendation services for users based on a certain type or several types of user context, but seldom focuses on the construction of user context models. This paper will focus on the construction of a user context model based on data mining technology.

III. CONSTRUCTION OF A USER CONTEXT MODEL

A. *Statistical Analysis of Context Factors*

Part of the data needed in this study was obtained by questionnaire survey, and the design of the questionnaire mainly included the content, questions and answers. We can quickly obtain coarse-grained user context information by questionnaire survey, which improves the accuracy of data crawled by a web crawler.

SPSS was used to conduct statistical analysis on multiple-choice questions or related questions to be studied, and the relationship and proportion of multiple related questions can be visually seen.

(1) From the analysis of multiple choice response rate of context, it can be seen that the response rates of location, weather and time are higher, which are 39.87%, 27.24% and 28.24% respectively. Therefore, the choice of location, weather and time is relatively high, which is generally recognized by investigators. This is shown in Figure 1.

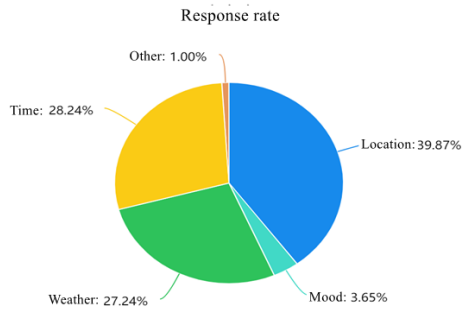

Figure 1. Pie chart of response rate

(2) As can be seen from the popularity rate of context factors multiple choice, the proportion of location, weather and time to the whole number of options is relatively high, which is 100%, 68.33% and 70.83% respectively. It can also be seen that location, weather and time are also the most recognized influencing factors. This is shown in Figure 2.

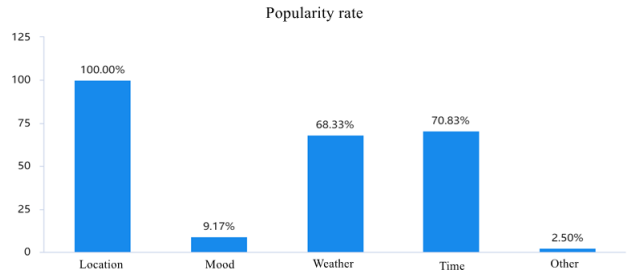

Figure 2. Line chart of popularity rate

(3) The Pareto Chart is the graphical embodiment of the 80-20 rule. 80% of the problems are caused by 20% of the causes. The cumulative ratio of location, time, and weather is less than 80%, so location, time, and weather are responsible for most of the problems. The cumulative ratio of mood and other options is much higher than 80%, so it isn't an influential factor in the main problem. This is shown in Figure 3.

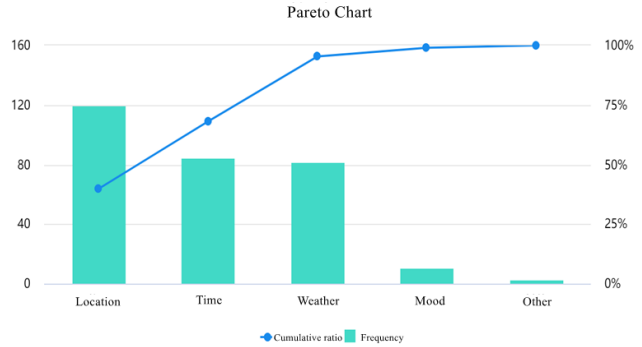

Figure 3. Pareto diagram

(4) The chi-square goodness of fit test was used to analyze whether the proportion distribution of multiple choices' options was uniform. It can be seen that goodness of fit test shows significance (χ^2=172.073, P =0.000<0.05), which means that the selection ratio of each item is significantly different, so the difference can be compared by response rate or popularity rate. Finally, to be specific, the response rate and popularity rate of location, weather and time are obviously the highest three items, so these are the three most influential context factors. This is shown in Table I.

TABLE I. SUMMARY TABLE OF RESPONSE RATE AND POPULARITY RATE

Item	Response		Popularity rate (*n*=120)
	n	Response rate	
Location	120	39.87%	100.00%
Mood	11	3.65%	9.17%
Weather	82	27.24%	68.33%
Time	85	28.24%	70.83%
Others	3	1.00%	2.50%
Summary	301	100%	250.83%
Goodness of fit test: χ^2=172.073 *p*=0.000			

(5) When the context factor of time is used as an option for multiple choices, there may be confusion between season and the morning, noon and evening of the day. So we should make a logical connection between a single choice and the time option in the multiple choice of context. When selecting the time, the single choice question "What time?" will appear. An illustration of the percentage of two choices can be seen in the cross plot. Only 9.14% of the people chose the season, while the 90.59% chose the morning, noon, and evening of the day. It can be seen that in terms of the context factors of time, season is not the main influential factor, while morning, noon and evening are the minor category that people care

most about under the category of time. This is shown in Figure 4.

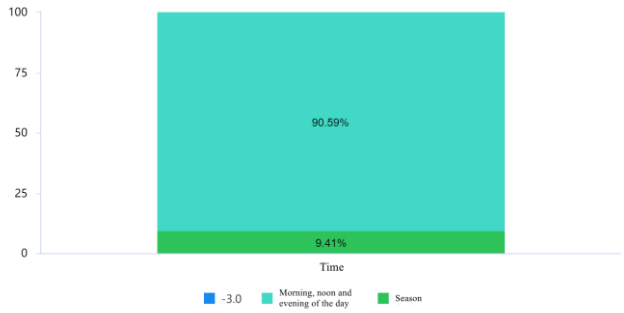

Figure 4. Cross plot

Through the above analysis of the questionnaire's questions and the related questions, it can be seen that all users believe that their shopping choices will be affected by context, and in the location, time, weather, mood, and other factors, the vast majority of people think location, time, and weather are the most important context factors that affect their shopping choices. Under the category of time context, the cross plot obtained by SPSS shows that most people think the time in the morning, noon and evening is the most influential. To sum up, through questionnaires and SPSS analysis, it is concluded that the above influential factors are the main factors for data collection in the later period.

B. Process of Constructing a User Context Model

The construction of a user context model is mainly through the training of clustering algorithm based on the dataset of specified web pages crawled by a crawler to form the model of corresponding rules. The main steps are six, as shown in Figure 5.

Step 1 Through the analysis of questionnaires, collect online shopping information of users, including time, weather and geographical location, determine the website address of the dataset you are looking for, and record the URL of the website.

Step 2 The imported urllib.request library is used to crawl and download the files and data needed in the web page of information collection.

Step 3 Delete the missing values in the crawling dataset, such as the row of the missing values in the location, through the data cleaning process. Replace and change the format of original crawling millisecond data with the time function to_datetime () in the pandas library.

Step 4 Call to_excel function to save the dataset that has been data preprocessed in the form of excel file. Because the size of data is uncertain, a variable engine='openpyxl' is usually added to the function so that the dataset can break through the original limitations.

Step 5 Through clustering, the saved dataset is clustered with Python to obtain the relevant model.

Step 6 Analyze the constructed user context model and draw relevant conclusions.

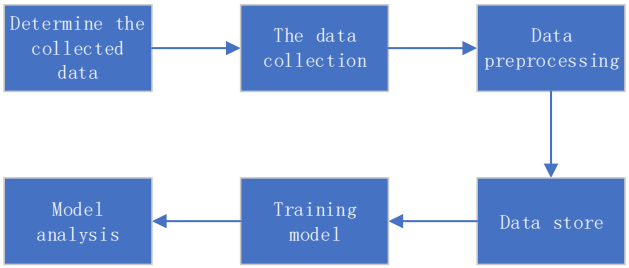

Figure 5. Process of constructing a user context model

IV. RESEARCH ON A USER CONTEXT MODEL BASED ON DATA MINING

A. Experimental Data

The experimental data in this paper were from the Tianchi data website of Aliyun, with a total of 70,000 pieces of original data. The data description is shown in Table II.

TABLE II. DATA DESCRIPTION

user_id	User identity, desensitization
itemid	Product ID, desensitization
behavior_type	User behavior types (including click, bookmark, add shopping cart and pay, represented by numbers 1, 2, 3 and 4 respectively)
user_geohash	Geo-location, security algorithm
item_category	Category ID (the category to which the product belongs)
time	The time at which the user action takes place
weather	Type of weather (sunny, overcoat, rainy, represented by numbers 1, 2, and 3 respectively)

B. Data Preprocessing

Due to the need of clustering, the original data set needs to be further processed to make it more consistent with the data standard of K-Means algorithm. Therefore, this paper changed all the data fields into numerical type through the previous process of data preprocessing. By deleting the line where the missing value was found, the final number of retained data was 60,395 pieces. The description of the new data fields is shown in Table III.

TABLE III. DESCRIPTION OF THE PROCESSED DATA

user_id	User identity, desensitization
itemid	Product ID, desensitization
behavior_type	User behavior types (including click, bookmark, add shopping cart and pay, represented by numbers 1, 2, 3 and 4 respectively)
user_geohash	Geo-location, security algorithm
item_category	Category ID (the category to which the product belongs)
time	Describe the time of day from 0 to 23
week	From 0 to 6 for Sunday to Saturday
weather	Type of weather (sunny, overcoat, rainy, represented by numbers 1, 2, and 3 respectively)

It can be seen from the above table that the original time has been changed into two new labels through format change, one recording the day of the week and the other recording the time of the day.

C. *Experimental Results and Analysis*

On the basis of experimental data, context and behavioral data of online shopping users were clustered, and the results of clustering were divided, and then the context model of users was depicted.

1) K value selection In the process of clustering experimental data, the K value in the experiment was determined by oneself, so the choice of K is different, which has a great influence on the clustering results. How to choose a good K value can be shown by the size of the contour coefficient. The specific code is as follows:

```
raw_data = pd.read_excel('D:/PC/workspace/data/user3.xls')
data1 = raw_data[['behavior_type','user_geohash','item_category','time','weather','week']]
from sklearn.metrics import silhouette_score
data2 = data1.sample(n=2000,random_state=123,axis=0)
silhouettescore=[]
for i in range(2,100):
    kmeans=KMeans(n_clusters=i,random_state=123).fit(data2)
    score=silhouette_score(data2,kmeans.labels_)
    silhouettescore.append(score)
plt.figure(figsize=(10,6))
plt.plot(range(2,100),silhouettescore,linewidth=1.5,linestyle='-')
plt.show()
```

Through the above code, the K-Means model of the data was iterated with the value of K from 2 to 100, and the contour coefficient can be used to select the appropriate number of clustering. According to the line chart, the point with the largest coefficient variation range can be found intuitively, and the point with the largest distortion range is considered to be the best number of clustering. This is shown in Figure 6.

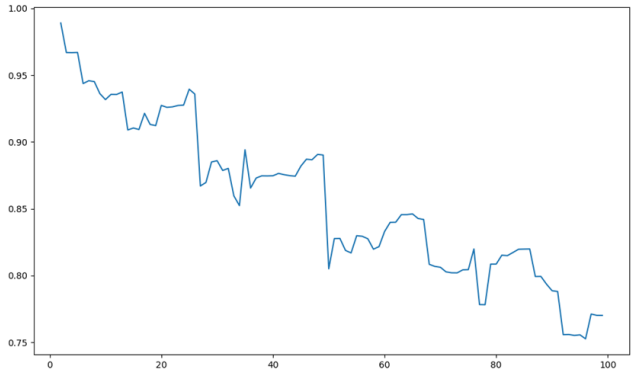

Figure 6. Line chart of contour coefficient

According to the description of contour coefficient line chart, it can be seen from the above figure that when K value is between 49 and 50, the distortion range of contour coefficient line is the largest. Therefore, it can be seen that the clustering effect is the best when K=50.

2) Analysis of clustering results By clustering the data, the results of clustering are shown in Table IV.

TABLE IV. PARTIAL CLUSTERING RESULTS

Category	behavior_type_x	user_geohash_x	item_category_x	time_x	weather_x	week_x	Number of categories
0	1.02	80495964314.06	10667.72	11.34	2.00	0.56	973
1	1.02	110050075643.89	11101.67	20.13	3.00	3.46	1251
2	1.02	104156482240.95	4396.95	9.23	1.00	2.11	1158
3	1.02	101526025178.99	4092.68	11.20	2.00	5.52	1113
4	1.02	70008870457.29	4110.42	14.58	1.00	2.51	1055
5	1.02	128552055287.48	2416.30	20.07	3.00	0.53	974
6	1.02	81687481355.95	4110.42	11.30	3.00	5.52	1072
7	1.02	81388622485.91	4119.75	15.28	1.00	4.89	1303
8	1.02	91402345398.82	10895.64	11.41	2.00	2.97	1435
9	1.02	114790691892.19	11212.64	20.50	1.00	2.46	1204
10	1.02	98694011841.89	4058.24	20.38	2.00	0.54	1731
11	1.02	75197927646.35	10889.87	10.56	3.00	2.89	1032
12	1.02	85788175379.03	4175.62	14.86	3.00	2.96	1314
13	1.02	112371007347.09	10999.98	20.11	1.00	5.15	1632
14	3.29	84902198698.65	6135.33	19.94	1.64	4.65	383
15	1.02	97741413551.43	10807.76	11.34	2.00	5.50	777
16	1.03	116194136992.88	3917.63	12.66	1.00	0.44	1316
…	…	…	…	…	…	…	…
38	1.02	88043081698.87	10745.59	12.16	3.00	5.15	1053
39	1.02	105155241587.27	4226.00	9.73	1.00	4.92	1085
40	1.02	76034775639.73	2246.66	20.49	2.00	2.93	1278
41	1.02	100724443311.10	6081.39	20.74	3.00	2.89	1556
42	3.31	95774616284.31	5606.19	11.45	2.55	1.54	388
43	3.31	113740974231.69	7352.23	16.74	3.00	4.74	328
44	1.02	108807620871.40	2214.68	20.71	1.00	2.93	1238
45	1.02	113828387899.41	6168.22	20.78	2.00	2.93	1579

46	1.02	95701809152.04	3925.92	20.36	1.00	0.54	1734
47	1.02	107979956569.45	4090.91	20.11	2.00	5.54	1676
48	1.03	118939435925.44	4269.01	14.98	2.00	2.97	1412
49	1.03	132414191623.62	2261.41	20.57	3.00	2.94	1293

Through the above table, the detailed situation of the user's context classification can be clearly got, for example, when the behavior_type_x is 1, user_geohash_x is 80495964314, item_category_x is 10667, time_x is 11, weather_x is 2, and week_x is 0, it means that in the area of geographical location 80495964314, the user clicks on the item in 10667 on a cloudy Sunday in the morning. Through such analysis, the clustering results of the collected user data can be described.

D. Analysis of Model Construction Results

The construction of the user context model is firstly obtained through questionnaires and the analysis of questionnaires, which is mainly affected by time, weather and geographical location. Secondly, through data collection of relevant factors and K-Means clustering operation of data, it is concluded that users are mainly affected by weekends and working days in the category of time context. Under the category of weather context, the main affected weather is sunny, cloudy and rainy. Due to the special encryption of the geographical location, only the encrypted area can be obtained, so the geographical location can only be expressed with encrypted numbers. So, the user context model is shown in Figure 7.

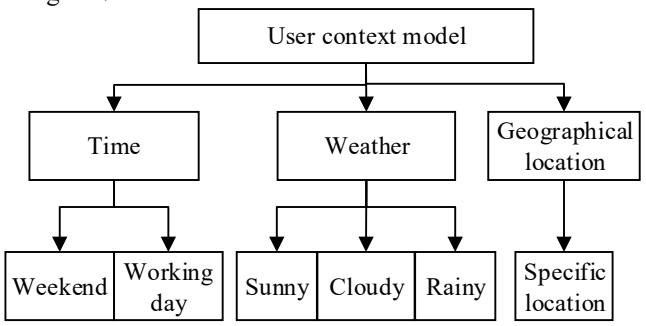

Figure 7. User context model results

V. CONCLUSION

Online shopping platform, as one of the platforms commonly used by users, needs to be able to better understand the needs of users and build a user context model to achieve the personalized needs of users. Through questionnaire survey and SPSS analysis, this paper made a statistical analysis on the context factors that may affect users' decisions. The user data that may influence the context factors of the user's decisions can be obtained by the web crawler, and the obtained data can be analyzed in fields. A better number of clustering was found through the contour coefficient, and the clustering model was obtained by using the obtained K value of the original data for clustering operation, thereby building the user context model.

There are some limitations in the model construction and analysis of this paper. The experimental data selected in this paper were the relevant dataset of Taobao users as an example, which cannot fully include all the sets of users in relevant context. In addition, only relevant time periods were selected as time periods of data collection in this paper, and the experimental data are not representative enough. In the following research, we will increase the experimental amount and acquire data in each time period to obtain more accurate user context data, build an accurate user context model, and put forward suggestions to improve user loyalty.

ACKNOWLEDGMENT

This work was supported by the Humanity and Social Science Youth foundation of Ministry of Education of China (18YJCZH077), the Social Development Program Project Supported by Science and Technology of Taizhou (TS202032), the "Qinglan Project" of Jiangsu Universities, University-Industry Collaborative Education Program (201802130070), and the Scientific Research Foundation of Taizhou University (QD2016036).

REFERENCES

[1] B. Wu, and Y. Ye, "BSPR: Basket-Sensitive Personalized Ranking for Product Recommendation," Information Sciences, 2020, 541.

[2] J. Zhang, C. Ma, C. Zhong, and L. Wang, "MBPI: Mixed behaviors and preference interaction for session-based recommendation," Applied Intelligence, 2021, (5), pp. 1-13.

[3] J. Liu, T. Zhou, and B. Wang, "Research progress of personalized recommendation system," Progress in natural science, 2009, 19(1), pp. 1-15.

[4] H. Kanh, and A. Wiener, "The Year 2000:A Framework for Speculation on the Next Thirty-Three Years," New York: MacMillan, 1967.

[5] J. Hong, E. H. Suh, J. Kim, and S. Kim, "Context-aware system for proactive personalized service based on context history," Expert Systems with Applications, 2009, 36 (4), pp. 7448-7457.

[6] D. Gavalas, M. A Kenteris, "Web-based pervasive recommendation system for mobile tourist guides," Personal and Ubiquitous Computing, 2011, 15 (7), pp. 1-12.

[7] S. Gómez, P. Zervas, D. G. Sampson, and R. Fabregat, "Context-aware adaptive and personalized mobile learning delivery supported by UoLmP," Journal of King Saud University-Computer and Information Sciences, 2014, 26(1), pp. 47-61.

[8] J. Hahn, "Indoor Positioning Services and Location-Based Recommendations," Library Technology Reports, 2017, 53 (1), pp. 9-16.

[9] W. Yue, Y. Wang, G. Wang, H. Wang, S. Dong, "Architecture of Intelligent Interaction Systems Based on Context Awareness," Journal of Computer-Aided Design & Computer Graphics, 2005(1), pp. 74-79.

[10] Y. Chen, Z. Liu, "Activity-based context awareness interaction design," Computer Engineering and Applications, 2013, 49(20), pp. 23-28.

[11] G. Ge, L.Yuan, X. Wang, "Personalized user interest modeling based on context aware," Application Research of Computers, 2017, 34(04), pp. 995-999.