A METHOD FOR TELECOM USER PORTRAIT MODELING

Tang Tingting[1,a], Yin Zhenyu[1,b] and Zou Yang[1,c]

[1] Chongqing University of Posts and Telecommunications, College of Computer Science and Technology, Chongqing 400065, China

[a]825993154@qq.com, [b]1461816223@qq.com, [c]zouyang@cqupt.edu.cn

Keywords: User portrait, K-means algorithm, User behavior

Abstract: In the era of big data, it is an important means to building user portraits and helping enterprises to implement precise marketing through comprehensive analysis of multidimensional data. Aiming at the problem of lacking detailed mining analysis and one-sided user attribute analysis, a method for modeling user portraits is proposed. On the basis of user's fact label, this method adopts the optimized K-means algorithm to extract the user's hidden label in order to fully describe the user behavior characteristics. The application results show that the modeling method can effectively extract the implicit information of users, fully reflect the potential demand of customers, and provide the possibility for accurate push marketing.

1. Introduction

With the large data technology in-depth research and extensive application, telecom companies are more and more focus on how to use large data to serve the precise marketing, so there will be a virtual representative of real user - user portrait[1]. User Portrait is a target user model based on a series of data which can be modeled according to the user's social attributes, consumer behavior, browsing behavior and other information to extract one or a class of user's label by structuring the user information. Through user portrait, we can understand the user, infer the user's potential needs, tap the potential user groups, and enhance the core of the enterprise influence.

In this paper, we proposed a method for modeling user portraits based on the optimized K-Means algorithm which can extract the user's latent label more effectively, characterize the user behavior and infer potential users' needs.

2. Related research

At present the user portrait has been applied in the industry. An J et al.[2] designed a method of creating user portraits based on real-time analysis of real social media data to provide accurate push information for online news media users. Zhang X et al.[3] proposed a quantitative bottom-up data-driven approach to creating user portraits which can better facilitate the user experience the research team to understand the user's workflow. In terms of telecom users, Zhang K [4] proposed a general structure that integrates customer information and mobile Internet behavior to analyze user portraits, but it lacks detailed mining analysis on user attributes and stays on the user's fact label.

In general, the user portrait has become one of the most effective tools to help enterprises accurately identify and analyze the target user. However, the target users in different industries in different areas have a greater difference, so we need targeted user portrait.

3 A method for telecom user portrait modeling

This article formed user fact label through basic information, SMS sent and other data processing. On this basis, the improved K-means clustering algorithm is used for depth data fusion and cross analysis to mine user potential information, form user hidden labels, and get user portrait.

3.1 Extraction of fact labels

The modeling method is based on different types of data objects which are shown in Table 1.

Table 1 the basic description of the acquired user data

Dataset name	Dataset description
User_Net_Log	S_IP、AD、BEGIN_TIME、URL、REF、UA、D_IP、COOKIE
User_Calling_Info	LATN_ID、ACTIVE_NBR、PASSTIVE_NBR、BEGIN_TIME、END_TIME、AMOUNT_TIME、CALL_TYPE_ID、OP_TIME
User_SMS_Info	ACTIVE_NBR、PASSTIVE_NBR、SEND_PACKAGES、REC_PACKAGES、SEND_BYTES、REC_BYTES、SP、EVENT_TYPE_ID、OP_TIME
User_Info	USER_ID、ACCS_NBR、SEX、AGE、AREA_CODE、IN_DATE、OUT_DATE

(1) User_Net_Log

The data is a collection of user network behavior, and it records in the user's Web site browsing, searching, clicking and other user behavior trajectory. Save S_IP, REF, UA, D_IP, COOKIE and other data fields, leaving the AD field as the user unique identifier. According to the method of [5], the following information about the user's online log is obtained:

$\{AD_i, NEWS\_C_i, VIDEO\_C_i, GAME\_C_i, READ\_C_i, BUSINESS\_C_i, BLOG\_C_i, DOWNLOAD\_C_i\}$

(2) User_Calling_Info

The data is a collection of user call records, and ACTIVE_NBR renamed to AD as the same identifier for user storage. Extract the n rows of AD or P_N as i to form an 8*n matrix, defined as follows:

$$CA_i = \begin{bmatrix} L\_ID_1 & AD_1 & PN_1 & BT_1 & ET_1 & AT_1 & CT\_ID_1 & OT_1 \\ L\_ID_2 & AD_2 & PN_2 & BT_2 & ET_2 & AT_2 & CT\_ID_2 & OT_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ L\_ID_n & AD_n & PN_n & BT_n & ET_n & AT_n & CT\_ID_n & OT_n \end{bmatrix}$$

Then we can get the following facts about the user's calling log: $\{FCD_i, ICS_i, OCS_i, AAT_i, BHC_i, FTC_i\}$

Among them, $FCD_i = \min\{BT_1, BT_2, \cdots, BT_n\}$; $ICS_i = |A_i| = a, A_i = [AD_1, AD_2, \cdots AD_i, \cdots, AD_a], AD_i = i$

$OCS_i = n - ICS_i$; $AAT_i = \frac{1}{n}(AT_1 + AT_2 + \cdots + AT_n)$; $FTC_i = n - BHC_i$;

$BHC_i = \frac{b+c}{2}$, $b = |B_i|, B_i = [BT_1, BT_2, \cdots, BT_i, \cdots, BT_b], 7:00:00 \le BT_i \le 23:00:00$,

$c = |C_i|, C_i = [ET_1, ET_2, \cdots, ET_i, \cdots, ET_c], 7:00:00 \le ET_i \le 23:00:00$.

(3) USER_SMS_INFO

The data is a collection of data sent by the user. Extract the n rows with the AD value of i to form a 9*n matrix, defined as follows:

$$SA_i \begin{bmatrix} AD_1 & PN_1 & SP_1 & RP_1 & SB_1 & RB_1 & SP_1 & ET\_ID_1 & OT_1 \\ AD_2 & PN_2 & SP_2 & RP_2 & SB_2 & RB_2 & SP_2 & ET\_ID_2 & OT_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ AD_n & PN_n & SP_n & RP_n & SB_n & RB_n & SP_n & ET\_ID_n & OT_n \end{bmatrix}$$

Then we can get the following facts about the SMS data label: $\{SPS_i, RPS_i, SBS_i, RBS_i\}$.

Among them, $SPS_i = sum(SP_1, SP_2, \cdots, SP_n)$, $RPS_i = sum(RP_1, RP_2, \cdots, RP_n)$, $SBS_i = sum(SB_1, SB_2, \cdots, SB_n)$, $RBS_i = sum(RB_1, RB_2, \cdots, RB_n)$.

3.2 Extraction of hidden labels

The user's hidden labels are usually not directly obtained from the user's behavior data. In this section, we mainly optimize the shortcomings of the traditional K-means algorithm and propose a method of cutting the maximum distance according to density.

3.2.1 User individual behavior label modeling

The user's individual behavior label contains the number of visits and the weight. The weight value is calculated by the following two steps. For example, the user’s browsing data is shown in Table 2:

Table 2 user AD_i's Internet log browsing data

AD	NEWS_C	VIDEO_C	GAME_C	READ_C	BUSINESS_C	BLOG_C	DOWNLOAD_C
i	150	20	41	30	80	48	56

①According to the advice of the telecom business experts, each type of web page' initial weight μ is different. The value of the longitudinal equilibrium is determined by the μ value. Examples of news type, such as (1). $NEWS\_W^v$ on behalf of news type web browsing behavior longitudinal equilibrium value. $NEWS\_C$ represents the current number of the user browsing the news type web.

$$NEWS\_W_i^v = \mu * NEWS\_C_i \quad (1)$$

②According to the different values of different labels, horizontal equilibrium is the conversion of the longitudinal equilibrium in accordance with the proportion. In (2), $NEWS\_W$ on behalf of the user news type web browsing behavior weight, and sum represents the sum of longitudinal equilibrium values of the user seven category website browsing behavior.

$$NEWS\_W = \frac{NEWS\_W_i^v}{sum} * 100\% \quad (2)$$

$$sum = NEWS\_W_i^v + VIDEO\_W_i^v + GAME\_W_i^v + READ\_W_i^v + BUSINESS\_W_i^v + BLOG\_W_i^v + DOWNLOAD\_W_i^v$$

3.2.2 K-means algorithm improvement

Because the acquisition of hidden labels often requires data mining methods, K-means algorithm [6] is a familiar clustering algorithm for mining users' hidden labels. The initial clustering center and the number of clusters are random, which leads to the problem that the local optimal solution or the clustering result is unstable. For this reason, many scholars try to improve K-means algorithm from different angles. In [7], the idea of data segmentation is proposed to determine the initial clustering center. In [8], a method based on maximum distance aliasing is proposed.

In this paper, we first calculate the two sample points from the furthest distance in the sample dataset X, then calculate the cutting point according to the density and the number of clusters, and finally calculate the initial clustering center.

Definition 1: dataset: $X = \begin{Bmatrix} x_{11} & \cdots & x_{1l} & \cdots & x_{1n} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ x_{i1} & \cdots & x_{il} & \cdots & x_{in} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{Nl} & \cdots & x_{Nn} \end{Bmatrix} \quad i = 1 \cdots N, l = 1 \cdots n$

Among them, $X_i = (x_{i1}, x_{i2}, ..., x_{il}, ..., x_{in})$ is any data object in dataset X, and represents the i-th dataset, which is a row vector in X. n represents the dimension of the data object. N represents the number of data objects in X.

Definition 2: Euclidean distance[9]: $d(X_i, X_j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{in} - x_{jn})^2}$ (3)

In formula (3): X_i and X_j represent any two data objects in dataset X. n represents the dimension of the data object. d represents the Euclidean distance between two data objects。

Definition 3: average Euclidean distance: $d\_average(X_i) = \frac{1}{k}\sum_{m=1}^{k} d(X_i, X_i^m) \quad m = 1, \cdots, k$ (4)

In formula (4): X_i is any data object in dataset X. k represents the number of clusters. X_i^m represents one of the k nearest neighbors of object X_i. $d\_average$ represents the average of the data object X_i and its k nearest neighbor Euclidean distance.

Definition 4: projection vector: $\alpha = \left(X_{start}^{(n)}, \cdots X_i^{(n)}, X_{i+1}^{(n)}, \cdots, X_{end}^{(n)}\right)^T$ (5)

In formula (5): n represents the dimension of the data object. $X_i^{(n)}$ is the projection of object X_i in dimension n. Projection vector α is the result set in ascending order after the $X_i^{(n)}$ value of all the objects in the dataset X falls within the interval $\left[X_{start}^{(n)}, X_{end}^{(n)}\right]$.

Definition 5: cutting distance: $d_{cut}^{(n)} = \left\lfloor \frac{N^-}{k-1} \right\rfloor$ (6)

In formula (6): k represents the number of clusters. N^- represents the dimension of the projection vector α. The cutting distance represents the dimension after dividing the projection vector $k-1$.

Definition 6: Davies-Bouldin index [10]: $DBI = \frac{1}{k}\sum_{i=1}^{k} \max\left\{\frac{\Delta(S_a)+\Delta(S_b)}{\Delta(S_a,S_b)}\right\}$ (7)

$$\Delta(S_a) = \max_{X_i, X_j \in S_a}\left\{d(X_i, X_j)\right\},\ \Delta(S_a, S_b) = d\left(\frac{1}{|S_a|}\sum_{X_i \in S_a} X_i, \frac{1}{|S_b|}\sum_{X_j \in S_b} X_j\right) \quad (8)$$

In formula (7): $\Delta(S_a)$ represents the intra-class distance of class S_a. $\Delta(S_a, S_b)$ represents the distance between class S_a and class S_b. k represents the number of clusters. In formula (8), $d(X_i, X_j)$ represents the distance between two data objects X_i and X_j in class S_a. $|S_a|$ represents the number of data contained in class S_a.

The improved algorithm is described as follows:

Input: sample dataset X, maximum number of clusters $k_{\max}$

Output: clustering results

(1) For $k=2$ to $k=k_{\max}$ do;

(2) Use Eq. (3) to calculate the Euclidean distance between all the objects in the dataset X, and find the furthest two points as X_{start} and X_{end};

(3) Use Eq. (4) to calculate the average continental distance of X_{start}、X_{end}、X_{start}^{m} and X_{end}^{m}. If X_{start} or X_{end} satisfies the definition of an isolated point [11], then exclude the point and turn to (2);

(4) Use Eq. (5) to calculate the projection vector α. According to the cutting distance $d_{cut}^{(n)}$ calculated by the formula (6), α will be divided equally among the $k-1$ vectors, and we can get $k-2$ cutting points $X_{cut}^{(n)}$ and the initial clustering center set $\{X_{start}, X_{cut_1}, X_{cut_2}, \cdots, X_{cut_{k-2}}, X_{end}\}$;

(5) repeat;

(6) Each data object in the dataset X will be assigned to one of the k clustering centers by the minimum distance principle;

(7) Recalculate the cluster center X_c for each class;

(8) Until the criterion function is converged;

(9) Calculate DBI using equation (7); turn to (1)

(10) Choose the best value k_{opt} of k, so that the DBI indicators to achieve the best;

(11) Output k_{opt} clustering results.

3.2.3 Extraction of hidden labels

The user's hidden label contains not only the label content, but also the weight, that is, the credibility of the label. The extraction algorithm is as follows:

Input: sample dataset X, maximum number of clusters $k_{\max}$

Output: hidden labels $\{IL_i, \omega\}$

(1)According to section 2.2.1, the sample dataset X is modeled by user individual behavior and we can get the processed dataset X';

(2)According to section 2.2.2, cluster analysis of X' by K-means improved algorithm, and get k_{opt} clustering results;

(3)Analyze the clustering results and obtain k_{opt} behavior preference labels $IL_i = \left\{x \middle| x \in k_{opt}\right\}$;

(4)Use the formula (3) to calculate the hidden label weight $\omega = d(X_i, X_c)$, $X_i \in S_a$ And X_c is the cluster center of the class S_a;

(5)Output hidden labels $\{IL_i, \omega\}$.

4. Experiment

This experiment uses Windows 7 system, open source data mining tool Weka 3.8, Java programming language, and 8.0GB memory.

4.1 Analysis of clustering effect

Test data using Iris, Wine and Balance-scale three sets of datasets which are frequently used to test machine learning and data mining algorithms. In this paper, the three sets of datasets were tested five times and compared with the traditional K-means algorithm. As shown in Table 3:

Table 3 the experimental results of contrasting with traditional algorithms

Algorithm count		Dataset					
		Wine		Iris		Balance-scale	
		initial center	accuracy	initial center	accuracy	initial center	accuracy
Traditional algorithm	1	97,150,243	0.6372	49,56,19	0.6733	77,74,164	0.6458
	2	123,187,423	0.6214	51,121,65	0.6667	75.36.43	0.4826
	3	165.86.129	0.6152	52.129.96	0.5833	78,124,49	0.6134
	4	104,255,485	0.6565	54,118,89	0.5733	77,68,140	0.5633
	5	158,2352,287	0.5335	50,76,108	0.6333	70,72,17	0.5263
	—	—	—	—	—	—	—
average accuracy of Traditional		—	0.61276	—	0.62598	—	0.56628
accuarcy of improved		—	0.72153	—	0.85764	—	0.70158

In the experiment process, we can find that although the clustering numbers of the two algorithms are the same, the results of the traditional K-means algorithm are different from each other because of the randomness selected by the initial clustering center. But using the improved algorithm based on density and maximum distance, the initial clustering center does not change with the number of tests, and the accuracy rate is relatively high and stable, so we can see that the algorithm can achieve good results.

The improved K-means algorithm is used to cluster the mobile log records of the mobile Internet users. The experimental data is a one-month data of 30,000 mobile users for a telecommunication enterprise server. The results are shown in Table 4.

Table 4 user log data classification clustered by improved K-means algorithm

class	count	Game	News	Video	Read	Business	Blog	Download
C1	598	20	21	2	25	25	3	4
C2	1078	3	31	25	5	34	0	2
C3	65	8	10	6	6	12	8	50
C4	8652	5	37	6	6	35	7	4
C5	1254	20	15	14	12	18	11	10
C6	2765	2	15	4	6	56	13	4
C7	284	74	5	2	0	15	0	4
C8	4017	48	8	7	2	17	13	5

According to the classification of Table 4, we can analyze the results. For example, users who classified as C1 prefer to browse news sites, business sites, game sites and reading sites. The analysis of the classification results is the hidden label of the classified user.

4.2 User Portrait Example

On the basis of the factual section of section 4.1, using K-means algorithm to extract user hidden labels, user portrait can be obtained, as shown in Fig 1.

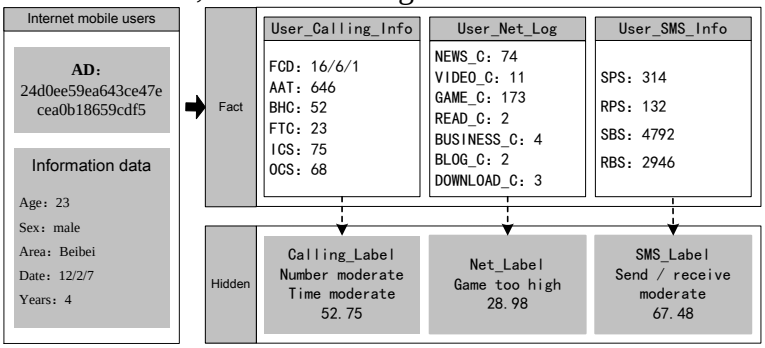

Fig 1 User Portrait Example

According to Fig 1, we can analyze that the user is in the middle level on the calling behavior and SMS behavior, but he is particularly interested in browsing the game site, so we can focus on pushing the game marketing information to the user.

5. Conclusions

This paper designs the telecom user portrait modeling method from the aspects of fact label and hidden label based on the user network log data, call record data, SMS sending data and information data. In addition, the self-improved K-means algorithm was used to mining user hidden labels and the visual view being combined with can clearly show the user portrait modeling process. As the user data obtained only covers the four aspects of the user, there are still some shortcomings in the hidden tagging of other users' behavior. The future paper will be focus on how user mobile attributes combined with network behavior to tap the user portrait and make it more fit with real users and better to provide theoretical and technical support for the telecommunications business.

Acknowledgments

This work was financially supported by Chongqing University of Posts and Telecommunications Doctoral Foundation (A2015-17).

References

[1] Shmueli-Scheuer M, Roitman H, Carmel D, et al. Extracting user profiles from large scale data[C]//Proceedings of the 2010 Workshop on Massive Data Analytics on the Cloud. ACM, 2010: 4.
[2] An J, Cho H, Kwak H, et al. Towards automatic persona generation using social media[C]//Future Internet of Things and Cloud Workshops (FiCloudW), IEEE International Conference on. IEEE, 2016: 206-211.
[3] Zhang X, Brown H F, Shankar A. Data-driven Personas: Constructing Archetypal Users with Clickstreams and User Telemetry[C]//Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. ACM, 2016: 5350-5359.
[4] Zhang K. Mobile phone user portraits in large data platform [J]. Information & Communications, 2014(2):266-267.
[5] Song S. Mobile Internet user clustering based on the analysis of URL [D]. Hebei University, 2013.
[6] Macqueen J. Some Methods for Classification and Analysis of MultiVariate Observations[C]// Proc. of, Berkeley Symposium on Mathematical Statistics and Probability. 2015:281-297.
[7] Zhu Y, Zhang C, Zhang B. Optimizing Research on K-means Based on Data Partition [J]. Computer Technology and Development, 2010, 20(11):130-132.
[8] Liu Z, Chen G. RFAT customer segmentation based on improved K-means algorithm [J]. Journal of Nanjing University of Science and Technology, 2014(4):531-536.
[9] Gower J C. Euclidean distance geometry[J]. Mathematical Scientist, 1982, 7(1):1-14.
[10] Davies D L, Bouldin D W. A Cluster Separation Measure[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 1979, 1(2):224-227.
[11] Angiulli F, Pizzuti C. Fast Outlier Detection in High Dimensional Spaces[C]// Principles of Data Mining and Knowledge Discovery, European Conference, Pkdd 2002, Helsinki, Finland, August 19-23, 2002, Proceedings. 2002:15-26.