Research on User Segmentation based on RFL Model and K-means Clustering Algorithm

Yunpeng Chen
National Library of Standards
China National Institute of Standardization
Beijing, China
chenyp@cnis.gov.cn

Yan Wang
Science and Engineering Department
Communication University of China
Beijing, China
wy@cuc.edu.cn

Ziyu Liu
Science and Engineering Department
Communication University of China
Beijing, China
caroline0539@cuc.edu.cn

Yao Qin
Science and Engineering Department
Communication University of China
Beijing, China
cuc_qinyao@sina.com

Abstract—With the rapidly shifting dynamics of the current market, the companies are seeking a more thorough method to research the preferences of their target market. As such, data mining models of user segmentation are often utilized to fill up the broadcasting and television research areas. This paper proposes a broadcasting and television RFL model for channel user segmentation and then gives the model for typical use case. The model has two main advantages, showing the users' value dynamically and having strong data availability together with wide model applicability. To define users' degree of satisfaction towards diverse television channel, R, F and L indicators are built. Then this paper uses the optimized k-means algorithm to divide users into clusters, along with cross validation by two-step clustering, which helps verify the results. By comparing each user cluster's average R, F and L indicators with the ensample mean, users can be subdivided into six levels: key-growth user, key-development user, general-growth user, key-kept user, low-value user and general user. On this basis, recommendations are given to the broadcasting and television operators and advertisers to assist them to make profits.

Keywords-User; Segmentation; RFL; k-Means; Clustering

I. INTRODUCTION

Given the rapid growth of the competitive market, the companies are trying to take advantage of its offerings and differentiate themselves from their competitors so as to be more competitive, which can be achieved by winning valuable customers. Thus, with the raising openness of the broadcasting and television market, establishing a marketing mechanism based on user segmentation is an inevitable trend.[1] To obtain more valuable users is the key to win a dominant position in the competition. To get more valuable users, a specific understanding of their needs is required. Thus, to understand such target market groups, it is necessary to segment the users.

Therefore, for broadcasting and television operators, analysis of users' viewing behaviors and scientific segmentations of users is the key to more market shares. As different people may watch a channel or a program for different reasons, user segmentation can be useful to identify homogenous groups of potential users and to develop customized strategies for each group [2], thus enabling companies to achieve the goal of profitability as well as customer satisfaction.

This paper proposed a modified version of RFM model as broadcasting and television RFL model for user clustering based on data mining, which focuses on the value of K -means clustering algorithm in the analysis of user value and provides channel user segmentation through analysis of the user's current and potential values. On this basis, enterprises may easily identify characteristics of various types of users in conjunction with channel features, which helps them implement differentiated user-oriented strategies and making appropriate decisions in marketing and advertising. [3]

The process of the research is mainly divided into the following four parts: the first part, using SAS statistical software for data preprocessing; the second part, building the broadcasting and television RFL model based on RFM model to calculate out the R, F and L indicators of each television channel; the third part, using K-means algorithm to partition users, who can be subdivided into six levels (key-growth user, key-development user, general-growth user, key-kept user, low-value user and general user) based on diverse characteristics of clusters; the fourth part, giving the RFL model for typical use case and discussing about recommendations and practical applications to broadcasting and television operators and advertisers.

II. THE DATA MINING MODEL OF BROADCASTING AND TELEVISION USER SEGMENTATION

A. *The RFL model*

The proposed RFL model is a modified version of RFM model (Table I), the latter is a method used for analyzing customer value, which is commonly used in database marketing and direct marketing especially in retail and professional services industries.

Based on RFM model, the established RFL model is proposed to define user's degree of satisfaction towards diverse television channel.

RFL stands for:

Recency - How recently did the audience watch?

Frequency - How often do they watch?

Length - How long do they watch?

To be more specific, researchers define

Recency = the number of days that have passed since the user last watched the television channel

Frequency = the times that the user watched the channel in the last few days

Length = time of the longest order from a given user

TABLE I. COMPARISON BETWEEN RFM MODEL AND RFL MODEL

Model	Recency	Frequency	Monetary/Length
RFM model	How recently did the customer purchase?	How often do they purchase?	How much do they spend?
Broadcasting and television RFL model	How recently did the audience watch?	How often do they watch?	How long do they watch?

The main idea of the RFL model is how to define the users viewing behaviors correctly so as to ensuring the consequence of user segmentation.

Statistics software SAS helps preprocess the origin viewing records of each sample user and calculate the R, F and L indicators towards diverse television channel.

Then the data mining technique k-means algorithm is used to create categories for each attribute. Each of the attributes has appropriate categories defined, and segments are created from the intersection of the values. If there were three categories for each attribute, then the resulting matrix would have twenty-seven possible combinations. In this model, a well-known commercial approach is applied by using five bins per attributes, which yields 125 segments as Fig. 1. On this basis, setting a score of five for the first twenty percent ordered users, a score of four for the next twenty percent…etc. Thus, each user can be put in a Cartesian coordinates, from (1,1,1) to (5,5,5). Users falling in an identical block are considered to have similar viewing behaviors. If a single indicator is required to evaluate all users, a product of three above-mentioned indicators will make it.

RFL model offers two main advantages. The primary one is showing the users value to the broadcasting and television operators to help them get more familiar with their users from the dynamic perspective, by bringing in the R, F and L indicators. What is more, the availability of data makes this model feasible. RFL model focuses on each user and consider R, F and L indicators of each instead of spending too much time observing user's specific viewing behaviors, which provides convenience and feasibility. Additionally, results are readily understood by business people in broadcasting and television industry. In the absence of other targeting techniques, it can provide a lift in response rates for promotions.

Plus, researchers collapse certain subsegments in this procedure, because the gradations appear too small to be useful.

If researchers divide users into 125 types according to the traditional RFM model, it will be hard to implement user-oriented service strategies. Therefore, further user segmentation turns out to be necessary. Researchers compare each user type's R, F and L indicators with the ensample mean, then label the user type with the up arrow when its mean is greater than the ensemble mean and otherwise, with the down arrow. Thus users fall broadly into six degrees (Table II). In some cases, certain user degree may be lost according to the television channel.

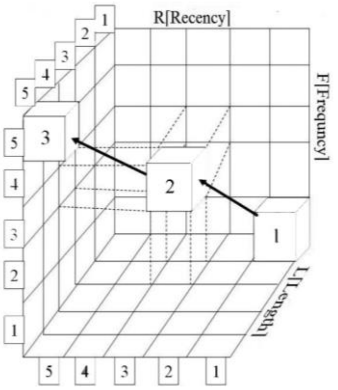

Figure 1. RFL model in three dimensional space coordinate.

TABLE II. USER DEGREE ON RFL MODEL

User type	RFL indicators	User degree
1	R ↓ F ↑ L ↑	Key-growth user
2	R ↓ F ↓ L ↑	Key-growth user
3	R ↓ F ↓ L ↓	Key-development user
4	R ↓ F ↑ L ↓	General-growth user
5	R ↑ F ↑ L ↑	Key-kept user
6	R ↑ F ↓ L ↓	Low-value user
7	R ↑ F ↑ L ↓	General user
8	R ↑ F ↓ L ↑	General user

The resulting segments can be ordered from most valuable (highest recency, frequency, and length) to least valuable (lowest recency, frequency, and length). Identifying the most valuable RFL segments can capitalize on chance relationships in the data used for this analysis. For this reason, researchers will use another set of data to validate the results of the RFL segmentation process in the following part.

B. K-means algorithm

K-means is one of the simplest unsupervised learning algorithms that solve the clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. Researchers usually place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early groupage is done. At this point researchers need to re-calculate k new centroids as barycenter of the clusters resulting from the previous step. After researchers have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result, it is clear that the k centroids move step by step until no more changes are done.

Finally, aiming at minimizing the objective function

$$E = \sum_{i=1}^{k} \sum_{p \in Ci} \left| p - m_i \right|^2 \tag{1}$$

As in (1), |p-mi|2 is a chosen distance measure between a data point p; and the cluster center mi.

The k-means algorithm is composed of the following steps:

a) Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.

b) Assign each object to the group that has the closest centroid.

c) When all objects have been assigned, recalculate the positions of the K centroids.

d) Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

C. *Distance cost function and the optimized k-means algorithm*

In cluster analysis, the k-means algorithm can be used to partition the input data set into k partitions (clusters). However, the pure k-means algorithm is not very flexible, and as such of limited use (except for when vector quantization as above is actually the desired use case). [4] In particular, the parameter k is known to be hard to choose when not given by external constraints, especially in spatial clustering. [5] For these use cases, many other algorithms have been developed since.

Thus researchers introduce a distance cost function and a new optimization algorithm [6] of k value to choose the optimization of k.

$$F(s,k) = L + D = \sum_{i=1}^{k} | m_i - m | + \sum_{i=1}^{k} \sum_{p \in C_i} | p - m_i | \tag{2}$$

In (2), m refers to the ensemble mean, mi refers to the sample mean of Ci, k refers to the number of clusters, p refers to any spatial objects, L refers to among-class distance and D refers to within-class distance.

When using distance cost function as spatial clustering validation function, researchers should follow the criterion of minimum cost [7], that is to say, when the distance reaches a minimum of the cost function, the spatial clustering result [8] will be optimal as well. Then the optimal k value is chosen by the following formula:

$$\min_k \{F(s,k)\}, k = 1,2,3,..., n \tag{3}$$

Based on k-means algorithm and the optimization of k value by distance cost function, the optimization k-means algorithm is composed of the following steps:

Input: the given Kr (Kr∈K) and spatial database.

Output: the optimized Kr* when a minimum of the cost function is reached.

a) Using k-means algorithm to realize the spatial clustering under all Kr value.

b) Calculating the distance value under different number of clusters based on the distance cost function

c) Finding the Kr* when the distance reaches a minimum of the cost function.

d) End.

III. EMPIRICAL ANALYSIS

This paper selects an area of about 50,000 actual users' viewing data for the study, collected from the radio and television set-top box in each user's house. Then researchers choose a television channel and give the RFL model for typical use case.

Firstly, calculating the R, F and L value based on the users viewing behaviors in a certain period. Then, convert the R, F and L value to standardized Z-score. (After Z-score conversion, the mean value is 0, and the variance is 1.) Setting k value equal to 4~10 for clustering, comparing the distance between the R, F and L center and the ensemble mean. According to the F-measure and Sig. value, it turns out that when k equals to 5, the final clustering center is as follows: (Table III).

TABLE III. THE RESULT OF CLUSTERING ON RFL INDICATORS

User type	RFL indicators			N
	R	F	L	
1	1.34018 ↑	0.46657 ↓	-0.37292 ↓	3080
2	-0.88258 ↓	4.62341 ↑	1.76295 ↑	201
3	-0.74523 ↓	1.43514 ↑	4.09324 ↑	313
4	-0.76186 ↓	1.00712 ↑	0.66104 ↑	1554
5	-0.51409 ↓	-0.30546 ↓	-0.30474 ↓	4937

Here are the interpretations of the above data: according to the positive or negative of the cluster center, users of the television channel is divided into three categories, which correspond to "low-value users", "key-growth users" and "key-development users" in the given six user types. (Table IV.)

The first category is user type 1, corresponding to the "low-value users": it has a total of 3080 users with following characteristics: long time having not watched the channel, or watching the channel in a low frequency and spending little time on it.

The second category is the user types 2, 3 and 4, corresponding to the "key-growth user": it has a total of 2068 users with following characteristics: watching the channel recently, watching the channel in a high frequency and spending long time on it.

Which, in turn it is subdivided into 3 subclasses:

Class1 (Type 2): it has total of 201 users, who watch the channel most frequently, are labeled "current users".

Class2 (Type 3): it has total of 313 users, who stay in the channel the longest, are labeled "willing to viewing user".

Class3 (type 4): it has total of 1554 users, whose viewing frequency and staying time are just slightly higher than the average, labeled "mediocre user".

The third category is the user type 5, corresponding to the "key-development user": it has a total of 4938 users with following characteristics: they watch the channel recently, but in a low frequency, and the time they stay in the channel is short. All types of users' distribution can be seen as Fig. 2.

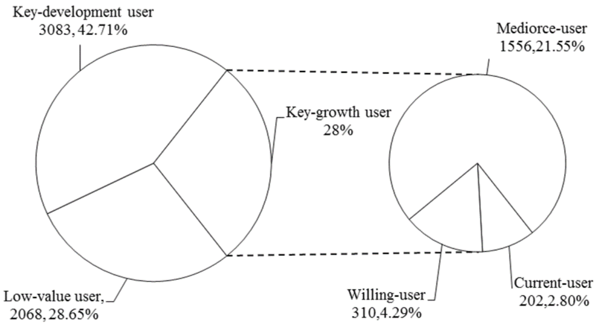

Figure 2. Users' distribution.

As for the above-mentioned channel, 15% of its users are "key-growth users", which refers to relatively loyal users. 51.8% are "key-development users" who can be transformed to loyal users by improving their viewing behaviors as RFL indicators. The rest 32.6% users are "low-value users", operators can give priority to giving up this part of the users in case of insufficient resources. Finally, a tag for each user of the channel is set according to the clustering results and users' characteristics [9].

In k means clustering, the set of k has a great effect on clustering, thus two-step cluster analysis method of hierarchical clustering algorithm is used to validate the k value. [10] It can be seen by the results of cross validation by two-step clustering that users of the channel are still divided into three categories as "low-value user", "key-growth user" and "key-development user". (Table IV.)

TABLE IV. THE RESULT OF CROSS VALIDATION BY TWO-STEP CLUSTERING

User type **User degree**	**1** **Low-value.**	**2** **Key-growth.**	**3** **Key-development.**
Size	39.2% (3952)	49.7% (5015)	11.1% (1119)
Features			
Zscore_R	1.09	-0.68	-0.77
Zscore_F	-0.46	-0.09	2.03
Zscore_L	-0.38	-0.16	2.06

TABLE V. THE RESULT OF CROSS TABLE VALIDATION

Two step clustering	**User type**			**Total**	**Overlap**
K means clustering	**1**	**2**	**3**		**ratio (%)**
1	3082	1	0	3080	99.97
2 (current)	0	202	0	201	100.00
2 (willing)	0	310	0	313	100.00
2 (mediocre)	7	606	943	1554	38.95
3	863	0	4072	4938	82.51
Total	3952	1119	5015	10086	

* User type 1 refers to "low-value user".
User type 2 refers to "key-growth user".
User type 3 refers to "key-development user".

As is seen from the cross table validation results, the user clustering results of two methods are basically the same, the difference in performance is, compared with the k means clustering analysis, 60.60% "key-growth users (mediocre)" is classified as key- development users and 17.49% "key-development users" is attributed to "low-value" users. It suggests that the result of k means clustering is effective.

IV. CONCLUSION

A data mining model of user segmentation fills up the broadcasting and television area. This paper proposes a modified version of RFM model as broadcasting and television RFL model, based on an optimization of k-means algorithm, for user segmentation and then gives the model for typical use case.

The RFL model offers six user degrees: Key-growth user, Key-development user, General-growth user, Key-kept user, Low-value user, General user. On this basis, recommendations are given to the broadcasting and television operators and advertisers:

For broadcasting and television operators, they can develop and adjust their marketing strategies for users of diverse television channel. Among them, "key-growth users" is considered the most loyal users towards the channel; "key-development users" and "general-growth users" refer to those who are potential to be developed to devoted users; key-kept users are urgent to be retained, or the channel may lose them. For low-value user, operators can give priority to give up this part of the users in case of insufficient resources.

For advertisers, they can find out their effective and valuable users to choose the right types and strategies of advertising.[11] For example, for a channel consisted of "key-growth users", centralized advertisements, series advertisements and marketing advertisements are supposed to be presented to its users; for a channel composed of a large proportion of "key-development users", repeated advertisements and persuasion advertisements can be delivered; plus, for a channel mainly made up of "low-value users", reducing the amount of advertising in order to better have a better control of the costs and maximize benefits.

ACKNOWLEDGMENT

This paper is financially supported by the Mining Research and Empirical Analysis Based on Cited Standard Intelligence Project of Central Fundamental Research Project (252013Y-3059), Engineering Planning Project of Communication University of China (XNG1356), Engineering Planning Project of Communication University of China (XNG1412) and Outstanding Young Teacher Training Project of Communication University of China (YXJS201527).

REFERENCES

[1] Muammer Ozer. User segmentation of online music services using fuzzy clustering. Omega, Volume 29, Issue 2, April 2001, Pages 193–206

[2] Tu S, Lu C. Topic-Based User Segmentation for Online Advertising with Latent Dirichlet Allocation[J]. Advanced Data Mining & Applications, 2010, 6441(2):259-269.

[3] Xueqing G; Xinyu G; Rong Z; Xiaofeng H, et al, Search Behavior Based Latent Semantic User Segmentation for Advertising Targeting, Data Mining (ICDM), 2013 IEEE 13th International Conference on , vol., no., pp.211,220, 7-10 Dec. 2013

[4] Hong-Bo G U. Study and Application of k Value Optimization Based on the k-means Clustering Algorithm[J]. Natural Science Journal of Hainan University, 2009.

[5] Ostrovsky, R., Rabani, Y., Schulman, L. J. and Swamy, C. (2006). The Effectiveness of Lloyd-Type Methods for the k-Means Problem. Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06). IEEE. pp. 165–174.

[6] Yang S L, Yong-Sen L I, Xiao-Xuan H U, et al. Optimization Study on k Value of K-means Algorithm[J]. Systems Engineering-Theory & Practice, 2006, 26(2):97-101.

[7] Xiaoyun, Wang; Shujun, Lei, Improved fuzzy c-means algorithm based on minimum of distance cost function, E -Business and E -Government (ICEE), 2011 International Conference on , vol., no., pp.1,4, 6-8 May 2011

[8] J. Stephen Clark1,2, Lukas Cechura3 and David R. Thibodeau4. Simultaneous Estimation of Cost and Distance Function Share Equations. Canadian Journal of Agricultural Economics/Revue canadienne d'agroeconomie,Volume 61, Issue 4, pages 559–581, December 2013

[9] Entreprise management publishing house: market segmentation[M].

[10] Sheng L, Xu X. A method of telecom consumer market segmentation based on the RFM model,2006,38(5):758-760.DOI:10.3321/j.issn:0367-6234.2006.05.025.

[11] .Xiaoyu Z, Xiaoyuan H, Fuquan S, et al. An Optimization Model for Promotion Mix Strategy Based on RFM Analysis. 2005, 13(1): 60-64. DOI:10.3321/j.issn:1003-207X.2005.01.011.