Research of Users' VOD Viewing Behaviour Based on ARMA Model

Xi Jiang, Yan Wang, Jian-ping Chai
Department of Science and Engineering, Communication University of China, Beijing, 100024, China

Abstract—**High-definition TV and interactive services bring a new viewing experience for the audience, and services such as video-on-demand and TV of yesterday have began to occupy a certain part of the viewing market gradually. Based on actual viewing records, this paper compares the differences between users' viewing time on live TV, video-on demand service and TV of yesterday service and employs the ARMA model to analyze users' VOD viewing behaviour. The results show that the overall users' VOD viewing behaviour presents a certain periodicity and users are more likely to enjoy VOD service on weekends than workdays. And the time series of viewing length on VOD per day is short-time related, so previous week's data can be used to predict future. The results of this paper can help TV network operators make strategy for business development.**

Keywords-viewing behavior analysis; video-on-demand service; time series analysis; the ARMA model.

I. INTRODUCTION

High-definition TV and interactive services have bring audiences a new viewing experience with less restrictions of time. Under this circumstance, analysis of users' viewing behaviour plays an important role in helping network operators formulating the development strategy. Existing TV services include TV live viewing, video-on-demand (VOD) service, TV of yesterday as well as other TV applications. Specifically, TV live viewing is the traditional TV business, while VOD allow users to select programs they like, TV of yesterday allow users to look back on the TV programs which have been broadcasted during the past week. Related studies have showed that broadcasting time is one of the important factors affect users' viewing behaviour[1].

Users' viewing behaviour can be regarded as time-series from time dimension, so time series data mining methods can be used to discover the hidden information and knowledge from historical data. According to different research tasks, time series data mining including time trend analysis, similarity search of time series, pattern mining of time series, clustering and classification of time series, time series visualization and time series prediction[2]. Appling these methods to the field of viewing behaviour analysis, audience rating prediction is one of the most popular issues. Specifically, there are two kinds of predicting methods, including prediction based on classification method[3,4,5] and prediction based on time series[6,7], the former method focuses on finding out factors that may affect the ratings and predict the level(for example, high, middle or low) of ratings while the latter can offer a more detailed numerical result on condition that the sequence is stable in the long term.

II. MODEL AND DATA SOURCE

A. ARMA model

The most common models for stationary time series prediction include the Auto Regression (AR) model, the Moving Average (MA) model and the ARMA model.

For AR system, the response at time t is directly related to the response before and the disturbance at time t. It has the following structure[8]:

$$X_t = \varphi_0 + \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \ldots + \varphi_p X_{t-p} + \varepsilon_t \quad (1)$$

Where $\varphi_p \neq 0$, and $E(\varepsilon_t) = 0, Var(\varepsilon_t) = \delta_\varepsilon^{\ 2}$, and $E(\varepsilon_S \varepsilon_t) = 0$ when $\mathrm{s} \neq \mathrm{t}$, and $E(\mathrm{X}_S \varepsilon_t) = 0$ when $\mathrm{s} < \mathrm{t}$.

For MA system, the response at time t is directly related to the disturbance at time t and before t. It has the following structure[8]:

$$X_t = \mu + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \ldots - \theta_q \varepsilon_{t-q} \quad (2)$$

Where $\theta_q \neq 0$, and $E(\varepsilon_t) = 0, Var(\varepsilon_t) = \delta_\varepsilon^{\ 2}$,and $E(\varepsilon_S \varepsilon_t) = 0$ when $\mathrm{s} \neq \mathrm{t}$.

For ARMA system, the response at time t is directly related to both the response at time before and the disturbance at time t and before. It has the following structure[8]:

$$\begin{aligned} X_t &= \varphi_0 + \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \ldots + \varphi_p X_{t-p} + \varepsilon_t \\ &- \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \ldots - \theta_q \varepsilon_{t-q} \end{aligned} \quad (3)$$

Where $\varphi_p \neq 0, \theta_q \neq 0$, and $E(\varepsilon_t) = 0, Var(\varepsilon_t) = \delta_\varepsilon^{\ 2}$, and $E(\varepsilon_S \varepsilon_t) = 0$ when $\mathrm{s} \neq \mathrm{t}$, and $E(\mathrm{X}_S \varepsilon_t) = 0$ when $\mathrm{s} < \mathrm{t}$.

The general process of time series analysis is showed in Fig.1[8].

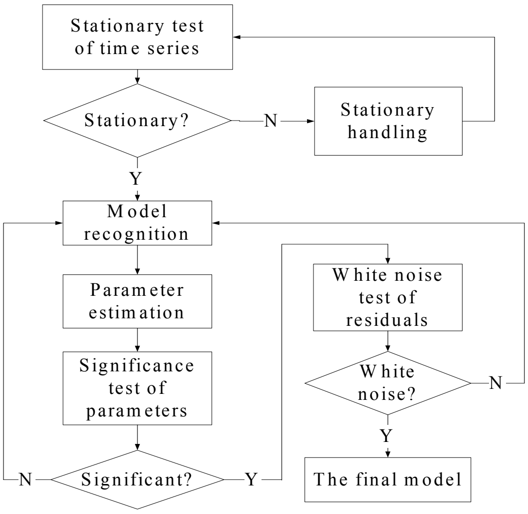


Figure 1. The general process of time series modeling

B. Data source

The viewing record used in this paper is provided by Beijing CUC-RZ Media Research Company. They are records of 125,456 families in Xi'an, Shanxi Province in 56 days, including 40 working days and 16 non-working days from March3,2014 to April27,2014. The format of viewing record is shown in Table1.

TABLE I FORMAT OF VIEWING RECORDS

User id	Date	Start time	End time	Channel id
$20.	Ymmdd10.	Time8.	Time8.	$8.

III. ANALYSIS OF USERS' VIEWING TIME

A. Overall viewing time analysis

In the 56 days, there are 125,456 viewing households, with 119,726 households (95.43%) viewing live TV, 62,053 households (49.46%) enjoying VOD service and 40,066 households (31.94%) enjoying TV of yesterday service. Calculating the number of online users for each hour of the day and computing an average. As shown in Fig.2, there are significantly more users watching live TV than VOD or TV of yesterday. For these three kinds of TV services, the maximum number of users appears at 9 p.m., 8p.m. and 10p.m. respectively, and the corresponding number is 33,547, 6,364 and 2,132.

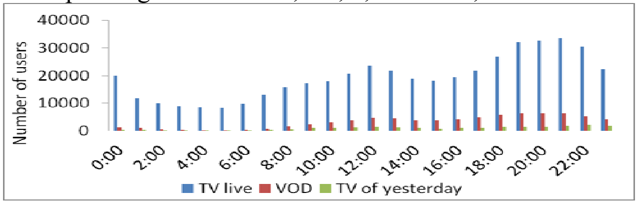


Figure 2. Number of viewing users at each hour of day

Calculate the overall viewing length spend on the three kinds of TV services and the result shows that TV live service occupies 86 percents of overall viewing length while VOD service accounts for 11 percent and TV of yesterday service accounts for 3 percent. Then calculate the proportion of viewing length spend on each service at each hour of day. As shown in Fig.3, the total viewing length of TV-on-demand and TV of yesterday occupy a quite small proportion from 0a.m. to 8a.m., the rate starts to grow from 8 a.m. to 5p.m. and achieves the maximum at 5 p.m. with a percentage of 17.75%, and then it declines a little and keep steady during night.

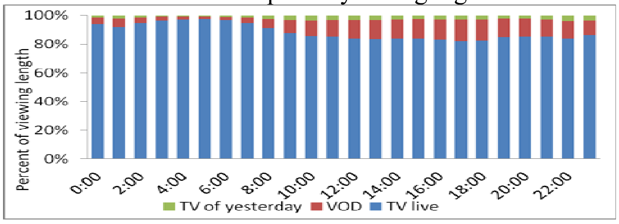


Figure 3. Proportion of viewing length of each TV service

As shown in Fig.4, the prime time of TV of yesterday service is later than that of VOD service. Specifically, the prime time of VOD service is from 6p.m. to 10p.m., with the maximum of viewing length arriving at 8 p.m.. And the prime time of TV of yesterday service is from 8p.m. to 11 p.m., with the maximum of viewing length arriving at 10 p.m.

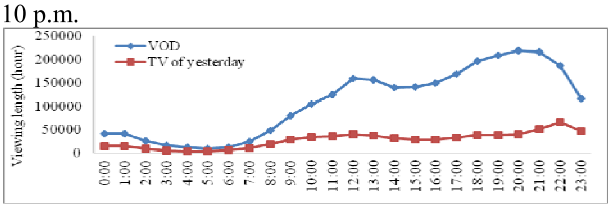


Figure 4. Viewing length spend on VOD and TV of yesterday

B. Analysis of VOD viewing

Calculate the viewing length on VOD service per day and the time series is shown in Fig.5. The sequence presents a certain periodicity and cycles for 7 days. As shown in Fig.5, there are 8 peaks and each peak corresponding to Saturday or Sunday, indicating that the users spend more time enjoying VOD on weekends than workdays.

Applying the stationary test of time series, calculating the autocorrelation and partial autocorrelation coefficient. As shown in Fig.6 and Fig.7, the autocorrelation coefficient fall outside two standard deviations when delay 1 order, 3 order,4 order and 7order, after 7order it's within two standard deviations. The partial autocorrelation coefficient fall outside the two standard deviations when delay 1 order, 2 order, 6 order and 7order, and after 7order it's within two standard deviations. Since autocorrelation function presents like tailing while partial autocorrelation function presents like truncation, indicating the AR(Atuo Regressive) model may be suitable. Applying the ADF(Augmented Dickey-Fuller) test and the result shows that the sequence does not contain a unit root and it's stable.

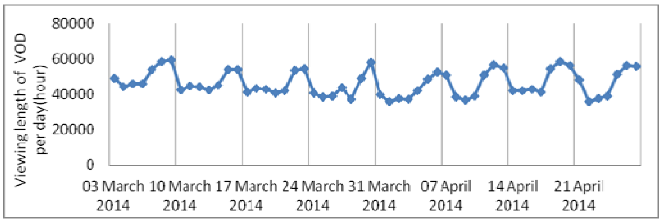
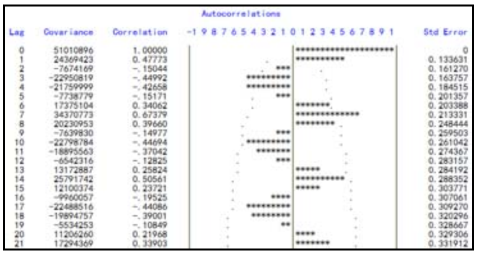


Figure 5. Time series of viewing length spend on VOD per day

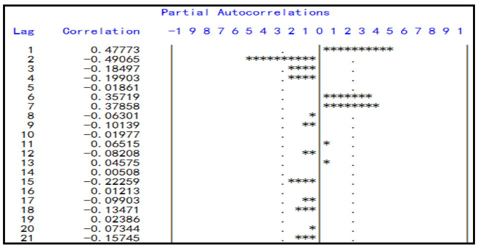


```
                                   Autocorrelations

Lag   Covariance   Correlation   -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1   Std Error

  0     51010896     1.00000    |                    |********************|          0
  1     24369423     0.47773    |                    |**********          |   0.133631
  2     -7674169     -.15044    |                 ***|                    |   0.161270
  3    -22950819     -.44992    |           *********|                    |   0.163757
  4    -21759999     -.42658    |           *********|                    |   0.184515
  5     -7738779     -.15171    |                 ***|                    |   0.201357
  6     17375104     0.34062    |                    |*******             |   0.203388
  7     34370773     0.67379    |                    |*************       |   0.213331
  8     20230953     0.39660    |                    |********            |   0.248444
  9     -7639830     -.14977    |                 ***|                    |   0.259503
 10    -22798784     -.44694    |           *********|                    |   0.261042
 11    -18895563     -.37042    |             *******|                    |   0.274367
 12     -6542316     -.12825    |                 ***|                    |   0.283157
 13     13172887     0.25824    |                    |*****               |   0.284192
 14     25791742     0.50561    |                    |**********          |   0.288352
 15     12100374     0.23721    |                    |*****               |   0.303771
 16     -9960057     -.19525    |                ****|                    |   0.307061
 17    -22488516     -.44086    |           *********|                    |   0.309270
 18    -19894757     -.39001    |            ********|                    |   0.320296
 19     -5534253     -.10849    |                  **|                    |   0.328667
 20     11206260     0.21968    |                    |****                |   0.329306
 21     17294369     0.33903    |                    |*******             |   0.331912
```

Figure 6. The autocorrelation function



```
                     Partial Autocorrelations

Lag   Correlation   -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1

  1     0.47773     |                    |**********          |
  2    -0.49065     |          **********|         .          |
  3    -0.18497     |               .****|         .          |
  4    -0.19903     |               .****|         .          |
  5    -0.01861     |               .    |         .          |
  6     0.35719     |               .    |*******             |
  7     0.37858     |               .    |********            |
  8    -0.06301     |               .   *|         .          |
  9    -0.10139     |               .  **|         .          |
 10    -0.01977     |               .    |         .          |
 11     0.06515     |               .    |*        .          |
 12    -0.08208     |               .  **|         .          |
 13     0.04575     |               .    |*        .          |
 14     0.00508     |               .    |         .          |
 15    -0.22259     |               .****|         .          |
 16     0.01213     |               .    |         .          |
 17    -0.09903     |               .  **|         .          |
 18    -0.13471     |               . ***|         .          |
 19     0.02386     |               .    |         .          |
 20    -0.07344     |               .   *|         .          |
 21    -0.15745     |               . ***|         .          |
```

Figure 7. The partial autocorrelation

Determine the order of model according to the Bayesian Information Criterion (BIC). The BIC value is minimum when p=7 and q=0, indicating AR(7) may be the right model. Estimate parameters using the least squares method, result shows that the parameter is significant at p=1 and p=7. Applying the white noise test to model with only $\varphi_7 \neq 0$ and model with $\varphi_7 \neq 0$ and $\varphi_1 \neq 0$, the result shows that the residual sequences of both models are white noise, indicating that both models are effective. Taking into account the simplicity of model, select the model with only $\varphi_7 \neq 0$ and the final model is showed below.

$$X_t = 50278.2 + 0.896 X_{t-7} + \varepsilon_t \tag{4}$$

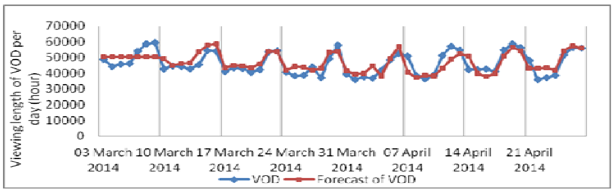


Figure 8. The AR model and the real time series

Figure 8. Shows the predicting time series based on the AR model and the real series of viewing length of VOD per day.

IV. Conclusions

Viewing behavior analysis of the specific region shows that live channels still occupy a major part of TV viewing market, with the percentage of 86% overall viewing length while VOD service accounts for 11 percent and TV of yesterday service accounts for 3 percent. The prime time of VOD service is from 6 p.m. to 10p.m., which nearly matches the traditional prime time of TV viewing (from 7 p.m. to 11 p.m.). And the prime time for TV of yesterday service is from 8p.m. to 11p.m., which is a little later than the traditional prime time. Users' VOD viewing behavior presents a certain periodicity and cycles for 7days. Users tend to spend to more time viewing VOD on weekends than workdays. The time series of VOD viewing length per day is short-time related, so previous week's data can be used to predict the future.

This paper focused on analyzing the overall users' VOD viewing behavior from time dimension, future study will center on applying the result to personalized service such as program recommendation.

V. Acknowledgements

This paper is financially supported by Engineering Planning Project of Communication University of China (XNG1356).

References

[1] Wang Bin-sheng. Critical factors of TV program ratings. *Journal of television engineering*, 2, pp.30-32 + 29, 2012.

[2] Jia Peng-tao, He Hua-can, Liu Li, Sun Tao. Overview of time series data mining. *Application Research of Computers*, 11, pp.15-18 + 29, 2007.

[3] Tu Juan-juan, Liu Tong-ming. A predicting model of TV audience rating based on the decision tree. *Micro Computer Information*, 27, pp.251-252, 2007.

[4] Zhang Jing, Bai Bing, Su Yong. Study of predicting TV audience rating based on the Bayesian network. *Science Technology and Engineering*, 19, pp.5099-5102, 2007.

[5] Chen Qing, Xue Hui-feng, Yan Li. Study on audio rating prediction based on semi-fuzzy kernel clustering algorithm. *Computer Engineering and Applications*, 48(6), pp.151-154, 2012.

[6] Liu Hui, Du Xiu-hua. TV ratings prediction method based on ARMA model. *Control Engineering*, 16, pp. 9-11, 2009.

[7] Yao Fang, Li Yue, The analysis of TV ratings of 30 TV channels based on time series. *Mathematics in Practice and Theory*, 13, pp.34-39, 2011.

[8] Wang Yan. Applied time series analysis. Beijing: Renmin University of China Press, 2008.