Box-Cox transformation as an alternative method for modeling video-on-demand popularity

María Teresa González Aparicio, R. García, Xabiel Garcia Pañeda, D. Melendi, S. Cabrero
Computer Science Department
University of Oviedo
Gijón, Asturias, Spain
{maytega, garciaroberto, xabiel, melendi, cabrerosergio}@uniovi.es

Abstract— **The popularity of multimedia videos related to a wide range of news, which were emitted in three different Spanish local on-line newspapers, has been researched in this paper. The statistic distribution from which the popularity came from is unknown. In fact, throughout the literature, many papers have modeled popularity with different distributions, such as Mandelbrot, Stretched, Zipf-like and so on. In this paper, the Box-Cox transformation has been proposed as a unified approach that would cover all the former distributions. The main advantage is its non-parametric nature and in consequence the model selection might be avoided.**

Keywords: Box-Cox, Mandelbrot, Stretched, Video-on-demand, Zipf-like.

I. INTRODUCTION

Nowadays the presence of streaming media on the Internet is becoming more popular, especially in web sites dedicated to news, sports, entertainment, education and even in the business world for marketing purposes. As a result, system designers have to face the new features of streaming media content, such as more computing power, an increase of bandwidth and storage requirements or a long-lived nature in order to supply good Web services [8]. Many technologies have emerged to manage this type of content and to reduce the impact over the different resources, among which could be mentioned multicast/unicast delivery, encoding formats or complex cache replacement policies, some of which are being improved steadily. However, more multimedia workloads have to be analyzed to achieve a well-known user access understanding.

In [5][10] an analysis of a video-on-demand service "La Nueva España" was presented, one of the services that is analyzed in this paper. Their studies highlight that content type, subjects, content update policy and even the content success make popularity a very difficult parameter to be modeled. A Zipf-like distribution has been applied in stable periods of time and an average θ was calculated. However, when the conditions of the service change due to the arrival of new content, a new value for θ is needed. An algorithm was defined but a popularity pattern was not established. Indeed, modeling user access is not an easy task, because there are so many variables involved. Accordingly, perhaps it is better to get rid of some of these variables and to start managing a simple service. For instance, the number of different types of contents on offer to the user could be reduced and focused to a specific topic and area. In this paper, we analyze session logs from three news video-on-demand streaming services, namely "La Opinión A Coruña" (www.laopinioncoruna.es), "Faro de Vigo" (www.farodevigo.es) and "La Nueva España" (www.lne.com). Each of them belongs to a different area of Spain.

As a result, we believe that our study provides relevant results for the design of news video-on-demand services. Specifically, it is focused on popularity distribution.

The rest of the paper is organized as follows. Section II reviews previous work. Section III presents a case study related to three news on-line video-on-demand services from Spain. An analysis of popularity with the three services has been carried out in Section IV. Finally, conclusions and future work are proposed in Section V.

II. RELATED WORK

The video access pattern has been analyzed in a wide range of media services (Web, file sharing, media broadcast, video-on-demand streaming). One of the first distributions applied to model access pattern was Zipf-like. In [4] a workload of one week was analyzed, with streaming-media sessions from 4,786 clients to 866 servers on the Internet, who accessed 23,738 different streaming-media objects. 78% were accessed only once, 1% were accessed ten or more times, and the 12 most popular objects were accessed more than 100 times each. The popularity distribution was modeled with Zipf-like with θ equal to 0.47. The conclusion was that accesses to streaming-media objects were less concentrated on the popular objects. Moreover, in [3] the behavior of the video access pattern was studied at different time scales (one month, six months and more than one year). Indeed, when the period was below seven months a Zipf-like approximation was possible with θ between 1.4 and 1.6, but not for longer periods.

In [7] sixteen workloads have been analyzed with different delivery methods (streaming, pseudo streaming, overlay multicast, P2P, etc), different sizes of media file, lengths of duration (from 5 days to more than two years) and different types of contents. The video access pattern has been fit with Stretched Exponential distribution despite of extraneous traffic, introduction of new content and recommendations [13], or "fetch-at-most-once" [2].

Many popularity distributions in streaming media show a circular curve on a log-log scale. Therefore, in [11] a Zipf-Mandelbrot distribution is proposed combined with a k-transformation to generate synthetic user access patterns. But, to date a non-parametric method has not been applied.

III. VIDEO-ON-DEMAND SERVICES

Principally, on-line newspapers give information with texts and images, and sometimes certain news items are combined with video. In this paper, we focus on video-on-demand services offered by three local on-line Spanish newspapers, which are called "La Opinión A Coruña", "Faro de Vigo" and "La Nueva España", located in three different areas of Spain, which are Coruña, Vigo and Asturias respectively. Their homepage has a link to a webpage, which offers news with videos, which are classified in several categories and subcategories. Some categories like "Sport" and "Society" are common to all three, but the rest are quite different. Moreover, each service offers some categories which refer to the area to which the service has a relation with. For instance, the category "Asturias" or "Aviles" is offered in the service "La Nueva España".

A. *Workload features*

These services have been studied during a period of nine months from 1st January to 30th September in 2009. Some services are more visited than others. Indeed, the service "La Nueva España" has 7% and 4% more visits than "La Opinión A Coruña" and "Faro de Vigo" respectively, and "Faro de Vigo" has barely 2% more than "La Opinión A Coruña". Taking into account that we are analyzing services focused on some areas of Spain, it is not unusual that one service has a higher degree of acceptance than other on-line news services among the population. Moreover, some services offer a higher number of news items than others. In Table I the number of new videos introduced to the system every day of the week during the whole period is depicted. The most remarkable service is "La Nueva España" where the number of new videos is four times higher than the other two services.

Table I. Number of new videos per service.

	"Faro de Vigo"	"La Opinión A Coruña"	"La Nueva España"
Monday	104	93	432
Tuesday	86	65	375
Wednesday	101	72	323
Thursday	35	39	348
Friday	64	90	368
Saturday	91	62	384
Sunday	100	48	390
Total	**581**	**469**	**2620**

Moreover, in Table I differences between the days of the week are shown. For instance, the number of new videos introduced on Thursday is lower than the rest of the days. It has to be mentioned that not all requests made every day were directed to new videos. Indeed, during the period studied, it can be observed that in "La Opinión A Coruña" the number of videos accessed was 163% higher than the number of new ones.

IV. DATA ANALYSIS

Data from accesses to multimedia videos on three on-line video-on-demand news services have been analyzed independently. The first six months have been used to find a unique model of the popularity for each service and the remaining period has been used to make a validation of the model.

Videos were sorted by their number of accesses in descending order, which is known as rank order distribution function in the social network. Modeling popularity has been studied throughout the literature with different distributions. Some of them have been considered in this paper, such as Mandelbrot, Stretched and Zipf-like. Due to the fact that each distribution has its specific parameters, the problem is to find the right values in order to provide a high quality fit to the data.

A. *Box-Cox Tranformation*

Box and Cox (1964) introduced a family of power transformation for a non-negative random dependent variable *y*. The method turns *y* into y(λ) where the family of transformations indexed by λ is expressed in the Equation (1).

$$y(\lambda) = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda} & si \quad \lambda \neq 0 \\ \log(y) & si \quad \lambda = 0 \end{cases} \qquad (1)$$

In the end, the transformation chosen gives the best λ in order to obtain the best fit to the data. The parameter λ is calculated with the maximum likelihood method [9]. Moreover, a confidence interval 100*(1- α)% is defined for λ [12], for a given significance level α, and it is expressed in the Equation (2), where $\hat{\lambda}$ represents the maximum value that λ could achieve.

$$\left\{ \lambda : \lambda > \hat{\lambda} - \frac{1}{2} \chi_1^{2*(1-\alpha)} \right\} \qquad (2)$$

The goal is to choose another possible value for λ inside this interval in order to generate an alternative model with an easy interpretation. For this reason, the most common values for λ might be {-2, -1, -1/2, 0, 1/2, 1, 2}. Consequently, if any of the common values for λ are included in the confidence interval defined, the common value should be chosen in its place. For instance, if λ=0.44 it would be hard to explain what the dependent variable y means, but $\sqrt{y}$ would be easier. In our study, the great majority of the confidence intervals for λ include the values 0 and -0.1. Therefore, any of them could be chosen as possible values for the Box-Cox transformation.

Box-Cox has two main features. On the one hand, it is a non-parametric method. On the other hand, it would offer a general linear model prediction such as Equation (3).

$$y(\lambda) = \beta_0 + \log(x)\beta_1 \quad with \quad 1 \le x \le n \qquad (3)$$

These features are quite important. Firstly, the statistical distribution from which the data would come is likely to be unknown. Secondly, it would be desirable to have a unified method to eliminate the problem of model selection. Finally, it

could allow the prediction of future demand based on independent past samples.

Following the procedure explained below, a unique linear model has been calculated for each service, and the final value for parameter λ was -0.1 for all cases. Consequently, our data was modeled with the expressions which are shown in Table II. Each predictor is presented with its standard deviation (s.d.) inside parenthesis. The variable x represents the ranking of the distinct multimedia videos.

Table II. Linear regression model with Box-Cox.

On-line news service	Linear model and R^2
"La Opinión A Coruña"	$y(\lambda)$ = -1.691209 (s.d. 0.4904398) * log(x) + 3.154359 (s.d. 0.731586) R^2 = 0.9702338 (s.d. 0.01772630)
"Faro de Vigo"	$y(\lambda)$ = -1.440526 (s.d. 0.590567) * log(x) + 2.862613 (s.d. 1.106257) R^2 = 0.9733878 (s.d. 0.01473696)
"La Nueva España"	$y(\lambda)$ = -1.838279 (s.d. 0.5783082) * log(x) + 3.450573 (s.d. 0.6787447) R^2 = 0.9730398 (s.d. 0.01341838)

The coefficient of determination (R^2) is quite high. However, if one sample is fitted with a linear regression and R^2 is high, it does not mean that the same linear regression model could be applied to another sample of the population, although it could be good enough for that sample. For that reason, the predictors β_0 and β_1 must have significance in the model. In other words, statistically they must be checked with a hypothesis test if $\beta_0 = \beta_1 = 0$ (p_value > 0.05), which we accordingly did [9].

In order to find a unique model several steps have been followed. Firstly, for each service and each day a regression linear model was calculated with Box-Cox. Secondly, for each model the significance of its predictors have been checked. Thirdly, Shapiro test has been used to check if the values of each predictor along the different samples follow a normal distribution [13], but it failed in all cases. Fourthly, the homoscedasticity per each predictor was tested with Levene's test, and it was successful in all cases. Fifthly, popularity was divided into groups by days of the week (from Monday to Sunday) and each predictor was compared with the F-Anova test [9], in order to detect any difference between the means of the seven groups. The F-Anova test was successful in all cases. As a result, the final value for each predictor will be the mean of all the means of the seven groups.

In Table III is presented the results for the p_value in the tests of Levene and F-Anova of one factor per each predictor for each group of days of the week are presented. Therefore, due to the fact that p_value > 0.05 in all cases, the test validation was positive as a whole. Following the same procedure, the values of the parameters for Mandelbrot [11], Stretched [7] and Zipf-like [2][3][5][10][13] distributions were calculated.

Table III. p_value of the predictors β_0 and β_1 per each model.

On-line news service	Predictor	p_value	
		Levene's test	F-Anova one factor
"La Opinión A Coruña"	β_0	0.8472	0.9555
	β_1	0.778	0.05
"Faro de Vigo"	β_0	0.6718	0.8536
	β_1	0.965	0.1741
"La Nueva España"	β_0	0.6558	0.151
	β_1	0.4026	0.2252

As previously mentioned, F-Anova test for one factor makes a comparison between the mean of each group, in our case each day of the week with the remaining days and a confidence interval is given per each pair of days. In Fig. 1 the results for the predictor β_1 in the service "La Nueva España" are given. Clearly, it can be observed that all confidence intervals include the value 0. As a result, there are no significance differences between the models of the different days in relation with the predictor β_1 and the mean of those values will be calculated as a representation of that predictor. The same has been done for the predictor β_0.

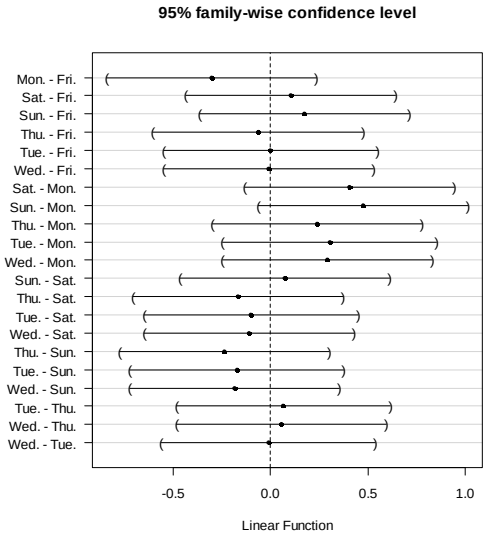

Figure 1. Confidence interval for predictor β_1 with F-Anova.

In the end, this procedure has allowed us to discover a general model for each day and for each service with the aim of being able to make popularity estimations for the different services. In Fig. 2 it can be observed how the general models that we have calculated for the Box-Cox, Mandelbrot, Stretched and Zipf-like distributions fit the popularity on 6th of September 2009 for the service "La Nueva España".

Actually, for this specific day, all of them seem to provide a high adjustment.

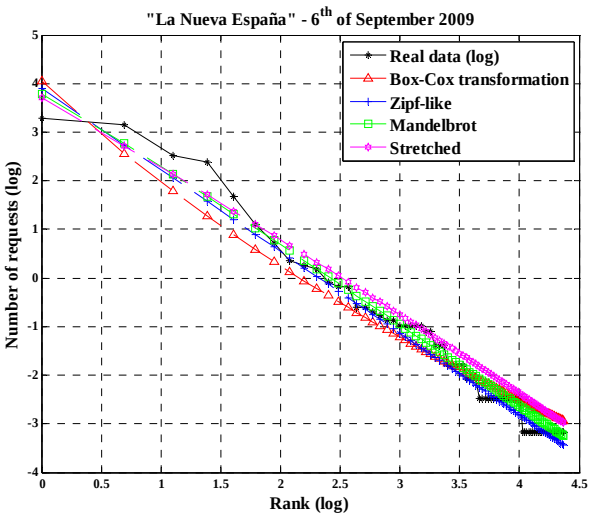

Figure 2. Fitting for popularity distribution.

A synthetic population was generated for the service "La Nueva España" with its linear model (Table II). Each day was fitted with the Box-Cox transformation. Once again, F-Anova test was applied to compare the value of the predictors between our synthetic population and the real ones. In Fig. 3 it can be observed that there are no significant differences in relation with the predictor β_1. Therefore, our model provides a good fit to the data and can generate reliable synthetic workloads.

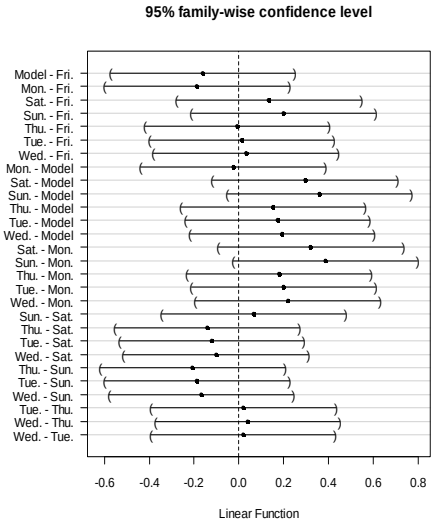

Figure 3. Confidence interval for predictor β_1 with F-Anova.

B. *Model Validation*

In order to evaluate the goodness-of-fit of the distributions Mandelbrot, Stretched, Zipf-like and Box-Cox transformation, the Kolmogorov-Smirnov (K-S) test has been used to decide if it is reasonable to assume if a sample comes from a population with a specific distribution. It is based on a comparison between the empirical cumulative distribution function (ECDF) and the theoretical one defined in Equation (4).

$$F(x) = \int_0^x f(y,\theta)dy \qquad (4)$$

The function probability distribution (pdf) is f(y, θ). Given n ordered data points x_1, x_2, ..., x_n the ECDF is defined as $F_n(x_i) = N(i)/n$, where N(i) is the number of points less than x_i (x is ordered from smallest to largest value). This is a step function that increases by 1/n at the value of each ordered data point. The test statistic used is expressed in Equation (5).

$$D_n = \sup_{1\le i\le n} |F(x_i) - F_n(x_i)| \qquad (5)$$

D_n is the upper extreme among absolute value differences between ECDF and theoretical CDF. This is checked with a hypothesis contrast where null hypothesis H_0 is if sample data come from the stated distribution and the alternative hypothesis H_1 is that sample data do not come from the stated distribution. The hypothesis regarding the distribution form is rejected if the p_value is lower than a significance level α. In this research the value chosen for α was 0.05.

Kolmogorov-Smirnov test has two important features. On the one hand, it can be applied when the sample size is not big enough. On the other hand, it does not need to bin the data. This is an important detail to take into account because the result could change depending on how the data is binned.

Many papers have used chi-square (χ^2) to measure the goodness-of-fit. The advantages mentioned before for the Kolmogorov-Smirnov test are the main drawbacks for χ^2 test. Indeed, in our study both methods have been applied. As a result, it could be proved that there were many cases where the χ^2 test passed when it should have failed. This phenomenon can be observed in Fig. 4. The p_value for K-S test was 0.005499604, but the p_value for χ^2 test was 0.9999887. In conclusion, the Kolmogorov-Smirnov test is more restricted and validation seems to be stronger.

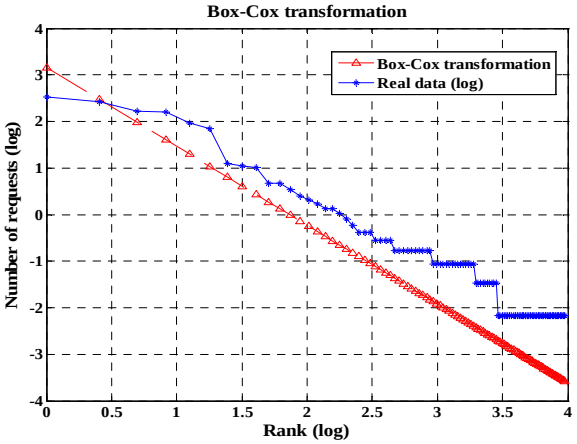

Figure 4. Fitting popularity with Box-Cox.

At the beginning, the method applied to obtain the value for the parameter λ followed the same process as the previous predictors β_0 and β_1. In fact, the F-Anova test was applied, but it did not pass the hypothesis (p_value < 0.05). So, it was considered to compare the median between the seven groups through the test of Kruskal-Wallis. The comparison was made with the value -0.1. In Table IV the results of the p_values are shown. Due to the fact that the p_values were higher than 0.05 the value of -0.1 was chosen.

Table IV. Test of Kruskal-Wallis for parameter λ.

	"La Opinion A Coruña"	"Faro de Vigo"	"La Nueva España"
Kruskal-Wallis	0.08403	0.2772	0.052

The model validation has been applied to each service every day during the three months from July to September in 2009. Therefore, the number of days for the validation period was 91 days in all. In Table V the number of times the K-S test had positive results (p_value > 0.05) is shown. As can be observed, Box-Cox transformation has a percentage of success slightly lower than the Stretched distribution in the service "La Nueva España". However, in general, the percentage of coverage during the whole period in the three services is higher in comparison with the rest of the distributions.

Table V. Number of success for K-S test (α = 0.05).

		News on-line service		
		"La Opinion A Coruña"	"Faro de Vigo"	"La Nueva España"
Models	**Box-Cox (λ = -0.1)**	34 (37%)	64 (70%)	78 (86%)
	Mandelbrot	18 (20%)	0 (0%)	0 (0%)
	Stretched	0 (0%)	4 (4%)	82 (90%)
	Zipf-like	24 (26%)	36 (40%)	53 (58%)

V. CONCLUSIONS AND FUTURE WORK

Many distributions have been researched in order to look for the best fit to the popularity. At the moment, Stretched distribution seems to be the best distribution which fits a wide variety of popularity [7] with different characteristics. The rest of the distributions could be applied when videos have specific features [10][11]. Although a specific workload might be fitted for a statistical distribution, the same type of distribution might fail to capture another workload [1]. For this reason, in our study, Box-Cox transformation is proposed as an alternative method for modeling popularity in news on-line video-on-demand. Moreover, it must be pointed out that a Zipf-like distribution is a particular case of Box-Cox (λ = 0).

Box-Cox provides good predictions with our data. It has to be taken into account that our videos have a short duration due to the fact that they are related to news. Therefore, a user is likely to watch the videos only once [2][6] due to their nature. Exceptionally, videos related to periodic events could be seen more times and follow a seasonal pattern [10]. Mainly, there are two important reasons why Box-Cox is a good choice. Firstly, it is a non-parametric model, therefore the unknown parameters could be determined from data. Consequently, data does not have to be fitted to any distribution. Secondly, this method provides a unified approach that covers all the lifetime distributions to avoid facing model selection.

Future work would follow two lines of research. On the one hand, a wider variety of news on-line video-on-demand services should be tested in order to confirm the Box-Cox applicability. Moreover, if longer periods of time [3][5] were modeled, the service administrator would schedule network resources and provide a higher quality of service for the users. On the other hand, the popularity of each video should be measured individually, and a classification should be carried out based on some common access pattern features [10]. Both lines will be combined in order to obtain a better understanding of video-on-demand behavior.

ACKNOWLEDGMENT

This work has been funded by different sources. On the one hand, the network operator Telecable from Asturias SAU and the Prensa Ibérica Editorial within the project MediaXXI(FUO-EM-174-07); on the other hand a National R+D Plan within the project TSI2007-60474. The authors are grateful for the data provided by the digital newspapers "La Opinion A Coruña", "Faro de Vigo" and "La Nueva España". This paper would not have been possible without their support.

REFERENCES

[1] Acharya, S., Smith, B., and Parne, P., "Characterizing User Access To Videos On The World Wide Web," *Multimedia Computing and Networking*, vol. 3969, pp. 130-141, January 2000.

[2] Cha, M., Kwak, H., Rodriguez, P., Ahn, Y., and Moon, S., "I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System," *USENIX/ACM SIGCOMM Internet Measurement Conference*, pp. 1-14, October 2007.

[3] Cherkasova, L., and Gupta, M., "Analysis of Enterprise Media Server Workloads: Access Patterns, Locality, Content Evolution, and Rates of Change," *IEEE/ACM Transactions on Networking*, vol. 12, no. 5, pp. 781-794, October 2004.

[4] Chesire, M., Wolman, A., Voelkert, G. M., and Levy, H. M., "Measurement and Analysis of a Streaming-Media Workload," *Symposium on Internet Technologies and Systems*, vol. 3, pp. 1-12, March 2001.

[5] García, R., Pañeda, X. G., García, V., Melendi, D., and Vilas, M., "Statistical characterization of a real video on demand service: User behaviour and streaming-media workload analysis," J*ournal of Simulation Modelling Practice and Theory*, vol. 15, pp. 672-689, 2007.

[6] Gummadi, K. P., Dunn, R. J., Saroiu, S., Gribble, S. D., Levy, H. M., and Zahorjan, J., "Measurement, Modeling, and Analysis of a Peer-to-Peer File-Sharing Workload," ACM *Symposium on Operating Systems Principles*, pp. 314-329, October 2003.

[7] Guo, L., Tan, E., Chen, S., Xiao, Z., and Zhang, X., "The Stretched Exponential Distribution of Internet Media Access Patterns," *Symposium on Principles of Distributed Computing*, pp. 283-294, August 2008.

[8] Kang, X., Zang, H., Jiang, G., Chen, H., Meng, X., Yoshihira, K., "Understanding Internet Video sharing site workload: A view from data center design," *Journal of Visual Communication and Image Representation*, vol. 21, no. 2, pp. 129-138, 2010.

[9] Myers, R. H., *Classical and Modern Regression with Applications*, 2d ed. United States of America, Duxbury Press, 1990.

[10] Pañeda, X. G., García, R., Melendi, D., Vilas, M., and García, V., "Popularity analysis of a video-on-demand service in a digital newspaper: influence of the subject, video characteristics and new content publication policy," *International Journal of Advanced Media and Communication*, vol. 1, no. 4, pp. 369-385, 2007.

[11] Tang, W., Fu, Y., Cherkasova, L., and Vahdat, A., "Modeling and generating realistic streaming media server workloads," *Journal of Computer Networks*, vol. 51, no. 1, pp. 336-356, 2007.

[12] Yang, Z., "Predicting a Future Lifetime through Box-Cox Transformation," *Journal of Lifetime Data Analysis*, vol. 5, no. 3, pp. 265-279, 1999.

[13] Yu, H., Zheng, D., Zhao, B. Y., and Zheng, W., "Understanding User Behavior in Large-Scale Video-on-Demand Systems," *European Conference on Computer Systems*, pp. 333-344, April 2006.