Analysis of the Correlation between User Behavior and User Engagement of Internet Video at Large-Scale

Yawei He[1, a *], Wenhui Zhang[2, b], Weili si[3, c] and Anming Wei[4, d]

[1, 2, 3] Communication University of China, Beijing, China

[4] Academy of Broadcasting Planning, State Administration of Press, Publication, Radio, Film and Television, Beijing, China

[a]qinkejing88@cuc.edu.cn, [b] zhangwenhui808@126.com,

[c]175599701 @qq.com, [d] weianming@abp2003.cn

Keywords: user behavior, user engagement, k-means, big data, Internet video

Abstract. As the development of the video over the Internet becomes rapidly and the user scale becomes continuously extend, user expectations and requirements for high quality of service are constantly increasing, which makes service providers focus on the quality of user experience (QoE). And how to assess QoE accurately and effectively become the first problem to solve. Considering that there are many factors affect QoE, this paper studies the correlation between user behavior and QoE. In this paper, by use the method of statistical analysis and data mining, we make a research on user logs of large scale in Internet video based on a data set captured from one of the largest video operators of China. We measure behavior metrics such as forward, replay, pause, full screen, and we use user engagement to quantify QoE at a per view level. The study shows that user behavior can reflect QoE, especially the number of fast forward and replay are significantly associated with QoE. Finally, this paper uses k-means clustering algorithm to divide users into three groups with different behavior patterns according to user behavior, which indicator that user with different behavior patterns have different engagement. And we also find that when the session include more than 10 times of fast-forward, the engagement in the session will decline.

Introduction

With the continuous development of the Internet technology and the increasing of network bandwidth, video distribution over the Internet goes mainstream and constitutes the main force of the Internet business. Given this context, it is important for service providers to focus on the evaluation of the quality of user experience, and it is crucial for service providers to improve QoE, as user expectations of Internet video service at large-scale are constantly increasing.

First of all, QoE is a measure of user satisfaction for the video service, which represent the integrated subjective feeling of user for the video service quality and performance. At present, academia and industry have carried out studies and made several conclusions on the research of QoE assessment and prediction fields. Review literature [1] summarized the influencing factors, quantitative methods and three kinds of evaluation methods of QoE, it consider the factors that impact QoE including three aspects: service, user and environment, and point out that the difficulty of QoE measurement lies in the impacts of user's subjective factors and environmental factors have on QoE are difficult to quantify. And at present, most of the research methods are stay in the service layer. As in [2], it study what impact of video quality metrics (such as: buffer, bit rate, etc.) have on user experience. And as in [3], it research the impact of metrics come from the network layer (such as: jitter, rate of packet loss, etc.) have on user experience. Although there are some research focused on the relationship between user engagement and attributes belong to user and environment aspects, as in [4], for different ways of video services(video-on-demand(VOD) or live), it proposed different user experience assessment models, but the results are less. So, based on this research background, we make a study to analysis the relationship between user behavior and user engagement.

In order to meet the growing customer demand of interactive request, video operators provide users with a lot of interactive interface, allowing users to occur a variety of interactive behavior according to their own needs, for example, user can decide the screen size and can control the playback progress according to their own preferences[5]. The generation of user interaction, not only meet the needs of users, but also provide us with the opportunity to have a glimpse of user experience, because there is a direct and indirect connection between user behavior and user experience. The direct relationships, such as: repeat viewing indicates that user experience is satisfying, and the indirect relationship, for example, pause can reduce the rate of buffering, and can affect QoE by increasing fluency of video.

Therefore, this paper uses the existing Internet video system to obtain user data record in a large number of network logs, and by the methods of statistical analysis and data mining algorithms, we analysis the relationship between user interactive behavior and user engagement. The user behavior in Internet video service we studies in this paper including: fast-forward, pause, replay, full-screen, and we use the fraction of the video viewed as a metric of user engagement. Our work mainly consists of two parts. First, we use statistical analysis methods to study the qualitative relationship between a certain type of interactive behavior and user engagement. The aim is to identify the behavior types that are closely associated with user engagement. Second, by the method of k-means clustering algorithm, we divide users into three different user groups with different behavior patterns according to behavior characteristics, and then investigate the distributions of user engagement in the three user groups. Our main observations are:

- User interactive behavior can reflect user engagement directly, especially the number of fast forward and replay have a strong relationship with user engagement.
- Replay is a positive behavior for engagement, and replay has the strongest correlation with user engagement compared to pause and full screen. And the number of fast forward is slightly negatively correlate with user engagement.
- According to user behavior, we can divide users into three groups with different user engagement and the session with replay and a less number of fast forward may have a high user engagement, while the session without replay but with a large number of fast forward may have a low user engagement.
- When the session include more than 10 times of fast-forward, user engagement will decline.

These results have significant implications on the development of a reasonable QoE prediction model for Internet Video. Considering user behaviors into QoE predict model can improve the reasonableness and accuracy of the measurement of QoE. And the results in clustering can provide support for video operators to divide users into groups according to behavior patterns, which can obtain user groups with different user engagement at the same time. Video operators can use user experience features to weaken the user's personality characteristics, so that to provide users with more targeted and group oriented services.

In this paper, Chapter 2 describe the preliminaries and datasets of the research, including data source and data preprocessing, and it introduce the metrics of user behavior and user engagement. Chapter 3 investigate the qualitative relationship between user behavior and user engagement, and Chapter 4 study the relationship between user behavior and user engagement by k-means, Chapter 6 is the conclusions and the future work.

Preliminaries and Datasets

User Behavior Metrics. The user behavior we studies in this paper is mainly user interaction, and the definition of the user interactive behavior is: the user under a particular viewing environment (e.g., computer, prime time, VOD), start a video session of a program, occurred a series of operations along the time axis. The interactive behavior generated by the user according to their own willingness. In our study, we consider four special interactive behaviors summarized below:

- Fast forward: Fast forward takes the largest proportion in user interaction, and user can control video playback process through seek operations (fast forward and rewind). On the one hand, as the Internet video system will mark out the video content now, user can jump to the interesting part directly through seek; on the other hand, user can seek to skip the boring content. If the user occur fast forward frequently, it may indicate that the user has little interest on the current video program, and when a few fast forward occurs, it may represent the user jumped to a certain point with the content of his or her interest. Therefore, we consider the number of fast forward operations as a behavior metric.

- Pause: Pause occur in the viewing process when user suffer buffer or due to their own reasons. User can pause to reduce the possibility of buffer, in order to get smoother viewing experience. The session with pause operation show that the user interest in the current video program is relatively high, and the user willing to sacrifice amount of wait time to continue watching the rest part of the content. But it does not means that the session without pause show a low user interest in the program, because user may successfully completed watching the program under good environmental conditions of network.
- Replay: Replay is a positive interaction for content providers, and the rate of replay also be used as an indicator to measure the popularity of the video program. When user finished their first viewing, user will chose replay for the reason of the content. Therefore, replay can indicate that user is interested in the current video content.
- Full screen: Internet video website using normal size of the window as the default, but allow user to switch from full screen, normal sized window and small screen according to their needs. In addition, only when user is interested in the video program, and user have full of confidence in the current network environment, will user chose full screen.

User Engagement Metric. In this paper, we choose user engagement as a metric for QoE, and engagement metric of user can be simply measured by session length, which reflect the involvement of user in a view session. The reason of this choice is that session length is an important parameter for advertisers to measure actual advertisement impressions, and user will end the session when the service is unacceptable for them, which makes session length become a metric reflect users' satisfaction. But considering the impact of video length for user engagement, we define user engagement as the fraction of the video viewed by user, and the calculation formula of user engagement metric we proposed in this paper is as follows:

$$UE = \frac{session\ length}{video\ length} \tag{1}$$

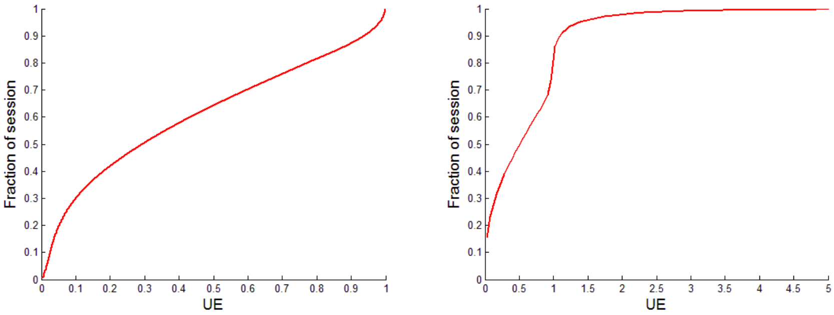
We use the fraction of the video viewed as user engagement metric is based on two reasons. First, this metric can be easily and objectively measured. Second, the calculation of user engagement metric takes into account the influence of video length. And we do acknowledge that this measurement has two potential limitations. First, we does not consider the situation that user playing the video in the background, which means the user have not watched the video actually, but might produce a session length with large value. Second, our study does not capture user attributes, such as user age, gender and job, these might have important impact on user engagement. However, our main purpose of this work is not to create a predictive model for user engagement, but to study the relationship between user behavior and user engagement.

Dataset and Preprocessing. The analysis of this paper is based on the data collected from one of the largest video operators of China. The format of the source data consist of two types of log: the access log and the player log, these logs record information such as: user IP, user interactive behavior and video quality metrics. Data covering 15days with more than 400 million users and more than 70 million video content involved, and the dataset including more than 800 million video sessions. Thus, the data we used are representative of the real situation in Internet video. In order to use the data efficiently, we collect all user behavior metrics described earlier as well as user engagement for each individual session. In addition we select session data come from VOD system with client of computer.

Correlation between User Behavior and User Engagement

In the study of the relationship between user behavior and user engagement, we use session as the basic unit, and we consider user behavior metrics including: the number of fast forward, pause, replay, full screen, and user engagement metric is the fraction of the video viewed in the session.

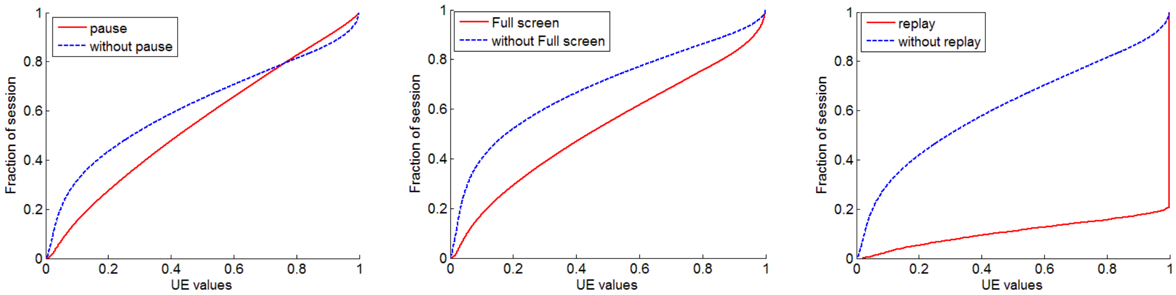
The distribution of user engagement. We begin with the distribution of user engagement on the dataset. Fig. 1 (a) is the cumulative probability distribution function (CDF) of user engagement, we can see user engagement include [0, 5]. Because we considered interactions such as: replay and pause, resulting in longer session length compared to video length, so the user engagement value is larger than 1. In order to facilitate the analysis, we limit user engagement value in the range of [0, 1], and the CDF of user engagement after processing is shown in Fig. 1 (b). About 20% of user engagement is less than 0.1, about 20% of user engagement is greater than 0.9, and about half of engagement are distributed in the ends.



(a) Original user engagement (b) Processed user engagement

Figure 1: The CDF of user engagement metric

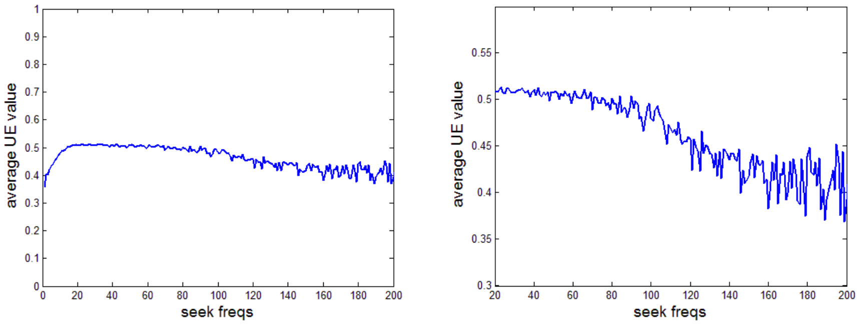
Relevance between user engagement and interactive behavior. In order to analysis the correlation between user behavior and user engagement, we separate video sessions into different subsets according to behavior, and then calculate the CDF of user engagement in each subset. For example, first, we divide the session process into two subsets named as "pause" and "without pause" according to whether a pause happened in the session; second, we calculate the CDF of user engagement in each subset, see Fig. 2 (a). Similarly, we calculate the CDF of user engagement in the subset divided according to replay and full screen, see Fig. 2 (b) (c). We can see that the distribution of user engagement in the two subsets exit difference regardless the type of interaction. But for replay, the difference of user engagement between the two subsets varies a lot, about 80% of user engagement metric value is 1 in the subset with replay, which means a high participation of user, while about 60% of user engagement metric value is less than 0.4 in the subset without replay. In contrast to pause and full screen, the difference of user engagement in the two subsets is small. These results show that replay has the strongest correlation with user engagement, and replay can reflect QoE, while pause and full screen may not.



(a) Pause (b) Full Screen (c) Replay

Figure 2: Qualitative relationships between user behavior metrics and user engagement metric

Relevance between user engagement and fast forward. Fig. 3 shows the relationship between the number of fast forward and user engagement. Each point on the curve is an average value of engagement metrics over the subset with the same number of seek operations during a session. Interestingly, we can see from Fig. 3 (a), the average value of user engagement increase with the increase of the number of fast forward when the number of fast forward is small, while when the number of fast forward is greater than 20 times, the average value of user engagement showing a decreasing trend with the increase of the number of fast forward, see Fig. 3(b). This may because frequent fast forward operations indicating that the user is not interested in the current video content, and certainly the performance of user engagement is low; but a few fast forward operations does not represent a low interest of user for the content, on the contrary, user can jump to the certain time point of interest directly to watch through a few times fast forward operations. Our analysis shows that the number of fast forward operations are more likely correlate with user interest, and the number of fast forward operations can reflect user engagement.



(a)The number of fast-forward operations (b) Part of (a)

Figure 3: Qualitative relationships between the number of fast forward and user engagement

In summary, we find that replay is a positive interactive behavior for user engagement, and replay has the strongest correlation with user engagement compared to pause and full screen. We also find that the number of fast forward is slightly negatively with user engagement, because frequent fast forward is more likely to improve the rate of buffering, and buffer will increase the session length. However, we cannot ignore the correlation between the number of fast forward and user engagement.

Clustering Analysis of K-Means

K-means clustering algorithm. K-Means clustering algorithm is a method based on distance, which need to set the number of clusters in advance. And K-Means belongs to the hard clustering, that is, any object must be divided into a category [6, 7]. The algorithm has good efficiency and the algorithm is simple and easy to implement, which is a common method of large data processing. Therefore, this paper uses K-Means clustering algorithm to dived users of Internet video into different user groups according to user behavior. The purpose of this paper is to obtain user groups with different user behavioral characteristics, and then investigate the characteristics of user engagement in different user groups, so that analysis the relationship between user behavior and user engagement.

In this paper, we divided user engagement into three categories of high, medium, low level, so we set the number of clusters to 3. By the method of K-Means clustering algorithm, the user interactive behavior is characterized as the input attributes in each session. According to the conclusion of the third part of this paper, we set up a two-dimensional as input attributes, including: the number of fast forward operations, whether the session contains replay behavior. And the output of the K-Means clustering algorithm is three collections of objects. The distance formula is Euclidean distance and the programming language is Python.

Vectorize user behavior. An important step of K-Means clustering algorithm is to describe the data objects in vectors. In this study, we take a large number of sessions as the research objects, and we regard every session as a data object in the clustering process. Since we investigate the user groups with similar behavior characteristics, we need to describe the user behavior in each session.

According to these regularities we concluded in the third part, we find out that the number of fast forward and replay behavior is strongly associated with user engagement. Therefore, we use the number of fast forward and replay behavior as the input to the K-means clustering algorithm, then each data object can be described as a two-dimensional vector. For example a session can be described as [0, 2], of which the first parameter indicates weather a replay occurred in the session. We require that if a replay occurred during a session, the first parameter record 0, otherwise referred to as 1. And the second parameter indicates the number of fast-forward operation occurs during the session. According to this rule, any session may be represented by a two-dimensional vector, denoted by [x, y], where X can take 0 or 1, while y take the natural number. Then the vector [0, 2] describes a session with 2 times fast forward, without replay behavior.

Results of k-means clustering. On the basis of the description rules we designed above, we divided user into 3 categories according to the user's behavior pattern by the method of K-mean clustering algorithm. And the clustering results are shown in Table 1.

Table 1. Results of K-means clustering algorithm

Content	**Cluster1**	**Cluster2**	**Cluster3**
Cluster center	(0.006,128)	(0.081,65)	(0.715,5)
Percentage	20.38%	11.17%	68.45%

We can see from the table, about 20.38% session belong to cluster1, and the cluster center is (0.006,128), the first metric means the average value of the number of fast forward is 128, and the second metric indicator that the possibility for the session in cluster1 containing replay is low. Cluster3 contains most of the objects, the percentage is about 68.45%, and the cluster center is (0.715, 5), a high possibility of containing a replay and the average value of the number of fast forward is small. And the cluster center of cluster2 is (0.081, 65), the percentage of cluster2 is about 11.17%. From the cluster center, we can see the effectiveness of K-mean clustering algorithm, it can be clearly seen that different clusters have different behavior characteristics.

Table 2 shows the statistical characteristics of user behavior and user engagement in the three user groups. We can see the maximum value of average user engagement appear in cluster3 with the maximum value of average replay frequency and the minimum value of average fast forward frequency. In addition, cluster1 has the minimum value of average user engagement, but the value of average replay frequency is small and the value of average fast forward frequency is high in cluster1. These findings indicator that we can divide user into three groups with different user engagement according to user behavior, and the session with replay and a less number of fast forward may have a high user engagement, while the session without replay and a large number of fast forward may have a low user engagement. And the conclusion we obtained above is consistent with our conjecture.

Table 2. Characteristics in the three clusters

Content	**Cluster1**	**Cluster2**	**Cluster3**
Ave fast forward freqs	128	65	5
Ave replay freqs	0.006	0.081	0.715
Ave user engagement	0.486	0.547	0.799

In order to observe the behavior characteristics and the distribution of user engagement in the three clusters we obtained through k- means algorithm, we count the number of fast forward and user engagement metric in the three clusters, and the CDF diagram is shown in Fig. 4. We can find that the number of fast forward contained in the session in cluster3 is the least, and is concentrated in [0,10]; and the number of fast forward in cluster2 distribute on [50,100]; and the distribution in cluser1 is [100,200]. We also find that about 70% of user engagement in cluster3 is 1; and the curve of user engagement in cluster1 and cluster2 are similar, but we can still find that user engagement in cluster2 is slightly greater than that in cluster1, which further verify the above conclusion that user behavior can reflect user engagement and user with different behavior patterns have different user experience of service. In addition, by the method of machine learning, we find that when the session include more than 10 times of fast-forward, user engagement will decline.

Conclusions and the Future Work

Our research finds out that user behavior can reflect QoE, especially the number of fast forward operations and replay are significantly associated with QoE. To this end, we divide users into three groups with different user experiences according to user behavior patterns by the means of k-means clustering algorithm, these results indicate that user behavior should be a key indicator in the measurement of QoE, which can allow video operators to evaluate QoE accurately. And the statistical techniques we proposed can be more broadly apply to measurement studies dealing with larger datasets. Our possible directions of future work include finding other potential factors that may impact or associate with QoE, and further to develop a reasonable QoE prediction model for Internet Video.

References

[1] Chuang Lin, Jie Hu, Xiangzhen Kong. The overview of QoE models and evaluation methods. Computer Journal, 2012.

[2] Dobrian F, Sekar V, Awan A, et al. "Understanding the impact of video quality on user engagement," Communications of the Acm, 2011, 41(4), pp. 362-373.

[3] Shen Y, Liu Y, Liu Q, et al. A method of QoE evaluation for adaptive streaming based on bitrate distribution[C]. IEEE International Conference on Communications Workshops. IEEE, 2014:551 - 556.

[4] Balachandran A, Sekar V, Akella A, et al. Developing a predictive model of quality of experience for internet video. Acm Sigcomm Computer Communication Review, 2013, 43(4):339-350.

[5] Hongliang Yu , Dongdong Zheng, et al. "Understanding user behavior in large-scale video-on-demand systems," Proceedings of the 1st ACM SIGOPS/EuroSys European Conference on Computer Systems 2006, Leuven, Belgium, Apr. 2006, pp. 333-344.

[6] Hartigan J A, Wong M A. Algorithm AS 136: A k-means clustering algorithm [J]. Applied statistics, 1979, pp. 100-108.

[7] Kanungo T, Mount D M, Netanyahu N S, et al. An efficient k-means clustering algorithm: Analysis and implementation [J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2002, 24(7): 881-892.