Personalized User Engagement Modeling for Mobile Videos

Lin Yang[a], Mingxuan Yuan[b], Yanjiao Chen[c], Wei Wang[d], Qian Zhang[a,*], Jia Zeng[b]

[a]*Department of Computer Science and Engineering*
Hong Kong University of Science and Technology, Hong Kong
[b]*Huawei Noah's Ark Lab, Hong Kong*
[c]*State Key Lab of Software Engineering, Wuhan University*
[d]*School of Electronic Information and Communications*
Huazhong University of Science and Technology

Abstract

The ever-increasing mobile video services and users' demand for better video quality have boosted research into the video Quality-of-Experience. Recently, the concept of Quality-of-Experience has evolved to *Quality-of-Engagement*, a more actionable metric to evaluate users' engagement to the video services and directly relate to the service providers' revenue model. Existing works on user engagement mostly adopt uniform models to quantify the engagement level of all users, overlooking the essential distinction of individual users. In this paper, we first conduct a large-scale measurement study on a real-world data set to demonstrate the dramatic discrepancy in user engagement, which implies that a uniform model is not expressive enough to characterize the distinctive engagement pattern of each user. To address this problem, we propose PE, a personalized user engagement model for mobile videos, which, for the first time, addresses the user diversity in the engagement modeling. Evaluation results on a real-world data set show that our system significantly outperforms the uniform engagement models, with a 19.14% performance gain.

Keywords: User Engagement, User Modeling, Mobile Video

*Corresponding author. Tel.: +852 23588766
Email address: qianzh@cse.ust.hk (Qian Zhang)

1. Introduction

The increasing prevalence of mobile devices has triggered an exponential growth in mobile video services. It is estimated that, by the end of 2018, mobile video will account for over two-thirds of the world's mobile data traffic [1]. In the wake of the development of screen size and computation power of mobile devices, users have a higher demand on the viewing experience. To cater for such needs, it is essential to accurately assess video quality.

The assessment of video quality has been widely studied by the multimedia community for a long time. Pioneer researchers have tried to quantify and improve users' viewing experience by optimizing quality-of-service (QoS) parameters [2, 3, 4, 5]. Although such QoS parameters are objective and easy to measure, their relationships to users' viewing experience are hard to quantify. To evaluate video viewing experience from the user's perspective, the concept of Quality-of-Experience has been proposed. A plethora of works try to solicit users' opinion evaluation score by conducting subjective tests [6, 7, 8, 9]. However, such subjective tests inevitably involve lots of human participation, thus are often in small scale due to the high cost.

In recent years, the concept of Quality-of-Experience has involved to *Quality-of-Engagement*. The user engagement, compared with the subjective and hard-to-measure user perceptual experience, is a more actionable metric to evaluate user's satisfaction with the video service and directly related to the service providers' revenue model [10]. As various parties are involved in the video service ecosystem, the user engagement can be evaluated from different angles. As a pioneer, Dobrian *et al.* collected a large-scale data set via client-side instrumentation and investigated how the video quality parameters affect the user engagement from the content provider's perspective [11]. The authors in [12, 13] developed a decision-tree-based engagement model to quantify the relationship between video-delivering QoS parameters and user engagement, which can help the design of content providers. Also, the authors in [14] examined the causal relationship between video quality and viewer behavior from the perspective

of content delivery network (CDN) owner, while another study utilized massive network-provider-side data to measure the impact of network dynamics on users' engagement in mobile video services [15]. Generally, these works leverage the power of machine learning and big data to reveal the complicated relationship between user engagement and confounding factors. Nevertheless, all of the existing models are built upon the entire user data set, averaging the effect of confounding factors on all users. When applied to individual users, such a uniform model may fail to characterize the distinctive patterns of personal user engagement.

To investigate users' differences in their engagement patterns, we collect a large-scale video streaming data set from the core network of a tier-1 cellular network in China. We first study the impact of the downlink throughput, which is an important factor from the perspective of network provider, on the user engagement in mobile video services. The result indicates that the same factor may have distinctive effects on different users. To further investigate the effect of user diversity on engagement modeling, we employ a wildly-used machine learning algorithm, *i.e.*, gradient boosted regression tree (GBRT) [16] to build a uniform user engagement model with data of all users, and individual user engagement models for selected users. Comparing individual models with the uniform model, we find that the model parameters of a specific user are considerably different from those of other users, as well as the uniform model. This implies that a uniform model is insufficient to comprehensively characterize the engagement level of individual users.

To gain a more accurate and fine-grained insight into user engagement, we need a personalized user engagement model which can comprehensively capture the user diversity. To achieve such a goal, there are several challenges: (1) The data set consists of millions of users, and building personalized engagement models for such a large user population is quite difficult. (2) The number of videos watched by each user is rather small compared with the total number of videos in the data set, insulting in a highly sparse viewing record, which makes it hard to build accurate models for each user. (3) While soliciting

information from accessory data sources is a potential solution to the sparsity problem, seamless integration of the information from various data sources is a non-trivial problem.

To tackle the above challenges, in this paper, we propose PE, a personalized quality of user engagement model for mobile videos from the perspective of mobile network provider, which takes user diversity into account and thus can provide a more accurate and fine-grained modeling. PE collaboratively learns the individual model for each user via matrix factorization and exploits the side information from other data sources to alleviate the data sparsity problem. The evaluations on a real-world data set show that PE significantly outperforms state-of-the-art user engagement models with a 19.14% performance gain.

With our system, mobile network providers can gain a more accurate understanding of user engagement with their services. Such knowledge can help them better invest network resources and perform case-by-case optimization [10]. Moreover, though our current implementation serves the need of mobile network providers, PE can easily be extended to meet the requirement of other service providers, *e.g.*, video content provider and CDN owner.

Our key contributions lie in three aspects:

- Our experiment on a large-scale video streaming data set demonstrates a significant user diversity in user engagement, which implies that the uniform model is insufficient for accurate engagement modeling.
- To the best of our knowledge, we are the first to propose a personalized user engagement model for mobile videos from the perspective of mobile network operators. This model can comprehensively capture the dramatic user diversity and provide a more accurate assessment of user engagement.
- We collect a massive video-related data set from a tier-1 network operator in China and perform a thorough evaluation of our system. The experiment results indicate our system can bring a 19.14% performance gain with respect to state-of-the-art user engagement models.

The rest of this paper is organized as follows. Section 2 defines the problem scope and validates the user diversity on a real-world data set. Section 3 formulates the problem and introduces the architecture of our system and Section 4 discusses the design of our personalized user engagement model. The evaluation results are reported in Section 5, followed with a related work review in Section 6 and conclusion in Section 7.

2. Problem Definition

Existing user engagement models are built upon the entire user data set, averaging the effect of factors on all users [10]. However, as users' viewing behavior diversifies, we expect the effects of these confounding factors to be disparate for different users and such diversity would affect the modeling of user engagement. To validate this, two natural questions follow: **(1)** ***Is there diversity in user engagement patterns?*** **(2)** ***If yes, how does such user diversity affect the user engagement modeling?***

In this section, we first define the engagement metric, then provide answers to the above two questions with experiments on a real-world data set.

2.1. Quantifying User Engagement

To quantify user engagement from the perspective of mobile network provider, we collect a large-scale anonymized IP flow trace from a tier-1 cellular network provider in China [17], which contains information of more than 8 million users and covers a large metropolitan area in one of the biggest cities in China from August 1st, 2014 to September 2nd, 2014 (the city name is anonymized for privacy issues). Through filtering and combining raw IP flow traces and signaling messages, we can obtain a fine-grained view of all mobile video sessions.

From the perspective of mobile network operators, fewer abandoned video-sessions and more downloading traffic is desirable, since a higher data usage results in a higher profit according to most revenue models of mobile network providers. Therefore, the *download ratio* is often used as a metric to measure

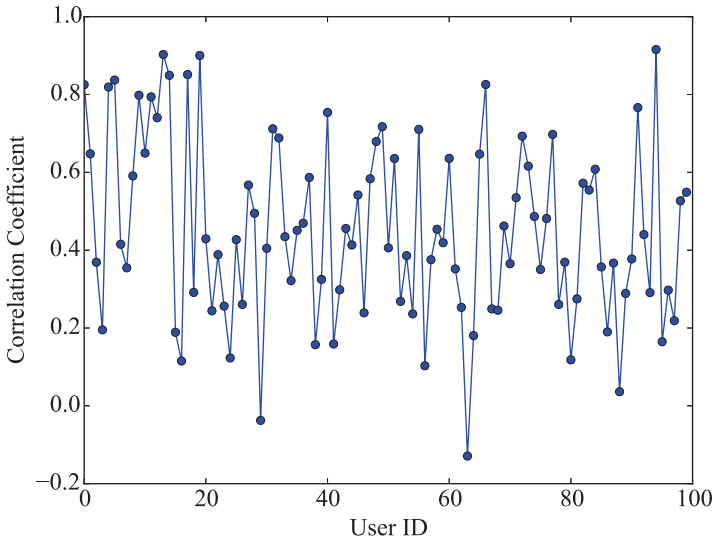

Figure 1: The correlation between video download ratio and average downlink throughput during the video session for 100 random users.

the user engagement from the perspective of network provider [15]:

$$\text{Download ratio } r = \frac{\text{downloaded bytes}}{\text{video file size in bytes}}. \tag{1}$$

In this paper, we will use this metric to quantify the user engagement, as it can be accurately measured from the network provider side. We understand that the download ratio can only capture the downloading phase of video streaming, but not users' behaviors after video download, *e.g.*, users may not watch the whole downloaded video due to lack of interest. Nevertheless, such events are out of the control of mobile network providers. Besides, other user engagement metrics suffer a similar problem, *e.g.*, video-played time can not reflect user engagement if the video is played in the background [12].

2.2. Validation of User Diversity

To find the answer to the first question on the existence of user diversity, we randomly select 100 users and compute the Pearson correlation coefficient between the download ratio and an important network quality indicator, i.e.,

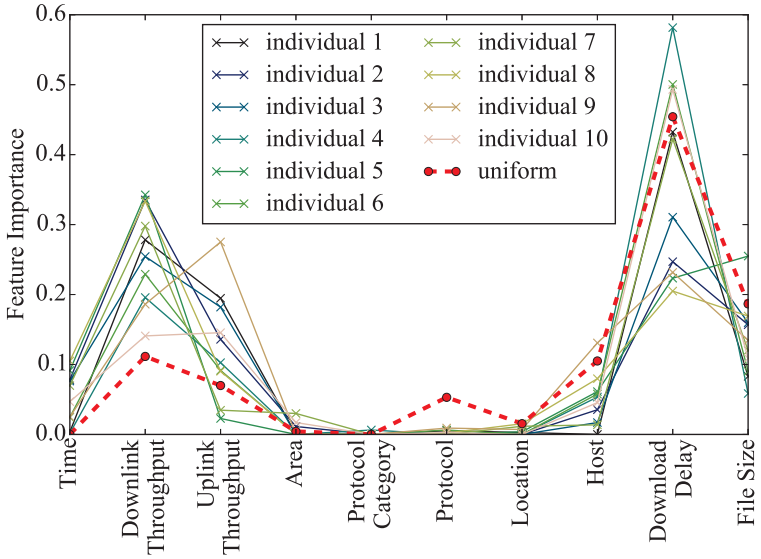

Figure 2: Feature importance of individual models and uniform model.

downlink throughput, for each user. Figure 1 shows that, the correlations between download ratio and downlink throughput change dramatically, spanning from -0.15 to 0.9. This variation indicates that the impact of downlink throughput on user engagement is quite diversified.

To further examine the second question of how such user diversity affects the engagement modeling, we adopt a mature machine learning algorithm, gradient boosted regression trees (GBRT) [16], to model the relationship between the user engagement and video-streaming-related features. We first feed all users' records into GBRT and build a uniform model. Then, we randomly select 10 users, each with a relatively rich video viewing record, and train individual models for each user. The comparison of the individual models and the uniform model is reported in Figure 2. Two interesting observations can be made. The first is that there exists a significant diversity in the feature importance of these 10 individual models, which corroborates our findings in the previous experiment on downlink throughput. For example, in the 8-th individual model, the feature importance of *download delay* is smaller than 0.2, but this feature

is weighted more than 0.5 in the 4-th individual model. Another observation is that the uniform model does not fit individual models well. The relative feature importance of the uniform model actually diverge from individual models, which implies that the uniform model only captures the average viewing patterns of users, but overlooks the diversity among individual users.

3. Design Overview

In this section, we introduce our data set, define the scope of our problem, then outline the architecture of our system.

3.1. Data Set

To build a reliable user engagement model, we have collected a large-scale data set from a tier-1 cellular network provider in China [17], which contains more than 8 million users and covers a large metropolitan area in one of the biggest cities in China from August 1st, 2014 to September 2nd, 2014.

This data set contains information from two data sources. On data source the raw IP flow trace captured from the links between the serving GPRS support nodes (SGSN) and the gateway GPRS support nodes (GGSN) in the core network of a 3G cellular network. It contains the flow-level information of all the IP traffic carried in the packet data protocol (PDP) context tunnels, that is, flows that are sent to and from mobile devices. This trace includes: start and end timestamps, anonymized user identifiers, traffic volume in terms of bytes, packet numbers for each flow, application information and location information. All user identifiers are anonymized to protect privacy without affecting our analysis. The other data source is the information from the user profile database, which is managed by cellular network operators to better understand users' needs. These information consists of user's demographic information, *e.g.*, age, gender, address, and data usage behavior, such as current data plan and data usage in the last month. To protect user privacy, all user-related identifiers are strictly anonymized and robust to de-anonymization.

3.2. Data Formulation

By extracting and aggregating these raw data traces (details of raw data processing are discussed in Section 3.3), we can obtain rich information about each video session, including the network statistics during the video streaming, the viewer demographic information, and their past viewing behavior patterns. We formulate the processed data as follows:

- m users, each of whom has l features. Let $D_{m\times l}$ denote the user feature matrix.
- n videos, each of which is associated with h video features. Let $S_{n\times h}$ denote the video feature matrix.
- $R_{m\times n}$ is the user-video matrix, in which r_{ij} is user i's download ratio for video j. $R_{m\times n}$ is a highly sparse matrix (sparse rate $\approx 99\%$). Since there are billions of videos and each user has only watched a tiny fraction of them, many items of r_{ij} are unknown.

In this context, modeling the user engagement is equivalent to building a model which can accurately predict the missing values in the user-video matrix R, based on the user feature matrix D and the video feature matrix S.

In the following subsection, we will discuss how to address the significant user diversity in the user-video matrix R, and exploit the information from the user feature matrix D and the video feature matrix S to alleviate the data sparsity problem.

3.3. System Architecture

In this work, we leverage a collaborative approach to build a personalized user engagement model, which is *de facto* a set of collaborative individual models. By collaboratively learning individual models for each user, this approach would make a better use of individual historical data and discover the latent connections hidden in users' viewing traces. To alleviate the data sparsity problem, we utilize the collective matrix factorization [18] to learn side information from the user feature matrix and the video feature matrix.

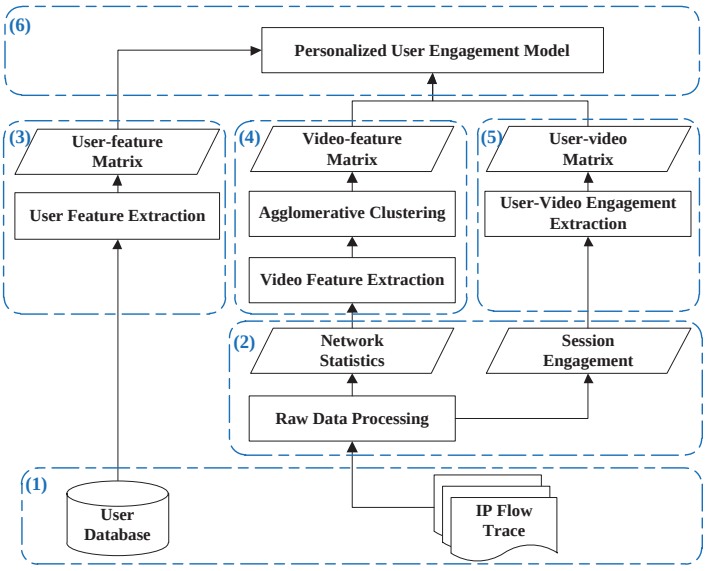

Figure 3: System overview of PE.

As shown in figure 3, our system comprises six major components:

(1) Data input. Our system mainly exploits two major data sets. One is the user profile database, which is operated by the network operator and includes rich user-side information. The other one is the IP flow traces collected from the core network of a 3G cellular network. It contains all the traffic traces at the IP layer, which can be used to extract video downloading records and the network quality during each video session. More information about this data set is presented in section 3.1.

(2) Raw data processing. Since the IP flow trace includes traffic records of all types of content, we first identify all video streaming flows by HTTP host and content-type headers, then aggregate them into sessions. For each video session, we extract and compute the corresponding network quality statistics during this session. Also, since we conduct engagement modeling from the network operator perspective, we compute the video download ratio as a metric

of user engagement.

(3) User feature extraction. In order to provide better services, network operators have collected rich user information, including personal profiles (*e.g.*, age, gender) and usage behaviors (*e.g.*, current data plan, bills and historical data usage). To protect user privacy, all user-related identifiers are strictly anonymized.

The user information can be very helpful for building a personalized model, but there is an abundance of user features, and we need to filter out "minor" features, which contain less information about users' preferences, to prevent the curse of dimensionality. This feature selection can be done via information gain analysis, which is a standard approach to uncover relationships between variables [19].

The underlying idea of information gain analysis is the entropy, which represents the informative level of a feature. The entropy of a random variable Y is defined as $I(Y) = -\sum_i P(Y = y_i) log P(Y = y_i)$, in which $P(Y = y_i)$ is the probability of $Y = y_i$, and the conditional entropy of Y given another random variable X, *i.e.*, $I(Y|X)$, can be computed as $\sum_j P(X = x_j) I(Y|X = x_j)$. The information gain then can be defined as $\frac{I(Y)-I(Y|X)}{I(Y)}$. To filter out features with less information, we can compute the information gain of each feature and use it as a measurement of feature importance. Due to page limitations, the details are omitted.

Through the feature selection, we select 24 out of 70 user features. These user features can be further categorized into two groups: (I) demographic information, which characterizes user population, *e.g.*, age, gender. (II) usage behavior, including current data plan, data traffic generated in last month, and so on. Table 1 gives some examples of user features.

(4) Video feature extraction. A major purpose of studying user engagement from the perspective of a network operator is to understand how it is affected by the network quality variation. Thus, apart from primitive attributes of videos (*e.g.*, file size, CDN server host), we also exploit the network quality statistics during a video session as video-session-associated features of this

Table 1: user & video features

Domain	Feature	Examples
User features	Demographic info.	Age Gender Living locale ...
	Usage behavior	Generated traffic in last month Current data plan Generated traffic in current month Streaming service accessing time ...
Video features	Video attributes	File size Video url CDN server host ...
	Session-associated features	Streaming protocol Round-trip time (RTT) Average download speed ...
	Context	Time Location (cell id)

video. We select 13 important network quality statistics via information gain analysis, and generally categorize them into three groups, *i.e.*, *video attributes*, *session-associated features* and *context*. Table 1 illustrates these video features.

A main concern of using the network quality as video feature is that a video can be streamed under various network qualities and results in multiple video-quality tuples in our data set, each of which corresponds to a specific network quality combinations. This will result in an explosive size of the video feature matrix. To reduce the computation complexity, we leverage agglomerative clus-

tering [20] to aggregate video-feature tuples that have the same URL and are streamed in a similar network quality condition into clusters. The number of clusters is a trade-off between the computation complexity and the granularity of network quality. After that, we merge videos that belong to the same cluster together and define the *video template* as the mean of all video-feature tuples in this cluster. Then, we represent the feature value of the video templates in a matrix format and incorporate it into our model.

(5) User-video engagement extraction. In our system, we quantify user engagement from the perspective of network operator via a continuous variable, the download ratio, which ranges from 0 to one to represent the fraction of video downloading. The data is transformed in a user-video matrix R, in which r_{ij} is the download ratio of the video j by user i.

(6) Personalized user engagement model. To address the user diversity, we choose to quantify the user engagement with a collaborative filtering model. Also, to alleviate the data sparsity problem, we employ a collective matrix factorization to integrate user- and video-feature data set. An in-depth discussion is given in section 4.

3.4. Limitations

We acknowledge that there are several potential limitations in our current system implementation:

- **Download ratio as user engagement.** Although the download ratio is directly related to the revenue model of mobile network provider, it constrains our analysis within the downloading phase. Some user behaviors after downloading, unobservable from the network side, can not be captured. However, as the download ratio can be accurately and objectively measured from network side and other analysis also employ the same metric [15], we use it as a start point for our analysis and our system can be easily applied to other metrics of engagement.

- **Data coverage over confounding factors.** As our data set is col-

lected from the network provider, several confounding factors that affect engagement are not captured in our data set (*e.g.*, video content and its popularity). As a result, our currently implementation of PE only provides a baseline performance and other data sources can be further integrated to provide a more comprehensive and accurate engagement assessment.

4. Personalized User Engagement Model

After data modeling, we have the user feature matrix D, video feature matrix S and user-video matrix R. Our goal is to build a personalized model to predict the missing values in R, with the help of D and S.

To build a personalized engagement model, one intuitive solution is to build an individual model for each user separately. However, this is impractical as there are millions of users. Another possible alternative is to first cluster users into groups, then build an independent model for each user group. This sounds like a reasonable solution, but it is based on a strong assumption that users with similar user features would behave analogously. This assumption poses a high requirement for the quality of user features. If the user feature does not fully capture the similarity of users on their engagement level, the clustering quality will be poor and eventually degrade the model's performance. Apart from that, this approach also understates the behavior patterns hidden in the historical data. For example, user i and user j are quite similar according to their user features, but indeed they behave in quite different ways (this may happen when the similarity of user features does not perfectly reflect user behaviors). In this case, even if there is adequate historical data records for both users, they will still be clustered into the same user group and thus share a comprised model which does not fit either of them.

To conquer the user diversity and data sparsity, we establish a model based on the collective matrix factorization framework [18]. The basic idea is that, we can first model each matrix via a low-rank approximation:

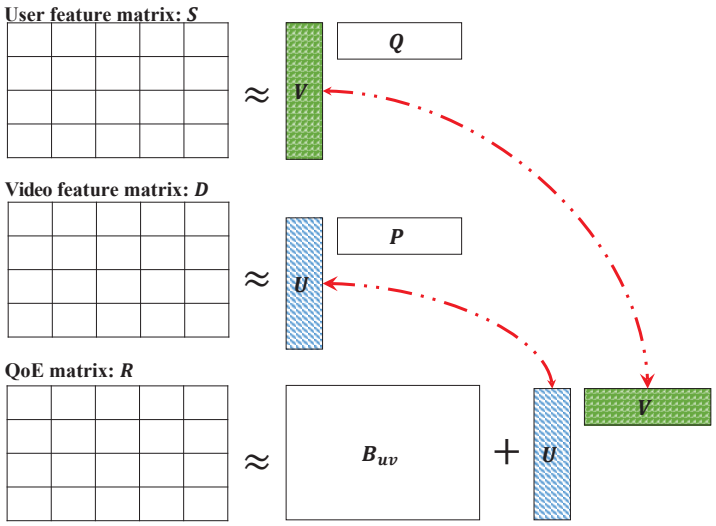

Figure 4: Overview of personalized user engagement model.

$$D = U \cdot P^T \tag{2}$$

$$S = V \cdot Q^T \tag{3}$$

$$R = U \cdot V^T + \underbrace{\mu J_{mn} + B_u \cdot J_n^T + J_m \cdot B_v^T}_{B_{uv}}, \tag{4}$$

where U, V, P and Q are latent factors, B_{uv} is the baseline predictor, μ is the average download ratio of all users, B_u, B_v are user bias and video bias matrices, and J_* are matrices of all the ones in different dimensions as suggested by their subscripts.

In this model, the user-video matrix R is approximated by the summation of baseline predictor B_{uv} and the product of latent factors U and V. The baseline predictor captures the basic engagement pattern and the bias introduced by each user and video, with which we can alleviate the cold start problem [21]. Meanwhile, the lower-rank approximation part, $U \cdot V^T$, characterizes the fluctuation caused by user and video diversity.

The rationale is that, by transforming both user and video to the same

latent factor space, we can estimate users' *interest* in these latent factors (*i.e.*, U) and the video's extent of these factors (*i.e.*, V). Each user i and each video j correspond to a preference vector $\vec{u_i}$ and score vector $\vec{v_j}$, and the learning procedure is conducted in a collaborative approach. We iteratively use all the video data of user i to help the training of user i's preference vector, and feed all data of users who watched video j into the model to learn the video score vector v_j.

This collaborative approach may still suffer from data sparsity problem as only users or videos which have common interactions (*i.e.*, users who have watched the same video, or videos that are watched by the same user) would collaborate. To further alleviate data sparsity, we also simultaneously factorize user feature matrix D and video feature matrix S and intentionally let latent factor U and V be shared among these factorizations. As a result, information from D and S can be propagated to R, and thus help gain a better performance. Figure 4 provides an intuitive overview of our model.

According to this model, we can formulate our objective function as:

$$
\begin{aligned}
&L(B_u, B_v, U, V, P, Q) = \\
&\parallel I_1 \circ (R - U \cdot V^T - \mu J_{mn} - B_u \cdot J_n^T - J_m \cdot B_v^T) \parallel_F^2 \\
&+ \frac{\alpha_1}{2} \parallel I_2 \circ (D - U \cdot P^T) \parallel_F^2 + \frac{\alpha_2}{2} \parallel I_3 \circ (S - V \cdot Q^T) \parallel_F^2 \\
&+ \frac{\lambda_1}{2} (\parallel U \parallel_F^2 + \parallel V \parallel_F^2) \\
&+ \frac{\lambda_2}{2} \parallel P \parallel_F^2 + \frac{\lambda_3}{2} \parallel Q \parallel_F^2 \\
&+ \frac{\lambda_4}{2} \parallel B_u \parallel_F^2 + \frac{\lambda_5}{2} \parallel B_v \parallel_F^2,
\end{aligned} \tag{5}
$$

where α_1 and α_2 are the reconstruction weights which control the degree of reconstruction and information sharing. The larger α is, the more important the corresponding term is in loss function and propagates more information to the others. $\lambda_i, i = 1, 2, ..., 5$ are the regularization parameters, and $I_i, i = 1, 2, 3$ are the indicator matrices where $I_{ij} = 0$ if the corresponding value is missing.

Let the operator $\circ$ denote the element-wise product of two matrices and $\|\cdot\|_F$ be the Frobenius norm.

In general, this objective function is not jointly convex, and we cannot get a close-form solution for minimization of this objective function. Therefore, we turn to search for a practical local optimal solution by gradient descent. More specifically, the gradients of loss function are:

$$\begin{cases} \nabla_{B_u} L &= E_r \cdot (-J_n) + \lambda_4 B_u \\ \nabla_{B_v} L &= E_r^T \cdot (-J_m) + \lambda_5 B_v \\ \nabla_U L &= E_r \cdot (-V) + \alpha_1 E_D \cdot (-P) + \lambda_1 U \\ \nabla_V L &= E_r^T \cdot (-U) + \alpha_2 E_S \cdot (-Q) + \lambda_1 V \\ \nabla_P L &= \alpha_1 E_d^T \cdot (-U) + \lambda_2 P \\ \nabla_Q L &= \alpha_2 E_s^T \cdot (-V) + \lambda_3 Q \end{cases} \tag{6}$$

where E_r, E_d and E_s are the residual errors with respect to R, D and S.

$$E_r = I_1 \circ (R - U \cdot V^T - \mu J_{mn} - B_u \cdot J_n^T - J_m \cdot B_v^T), \tag{7}$$

$$E_d = I_2 \circ (D - U \cdot P^T), \tag{8}$$

$$E_s = I_3 \circ (S - V \cdot Q^T). \tag{9}$$

With the gradients, we can resort to the gradient descent approach to iteratively minimize the objective function. The detail is described in Algorithm 1.

Although gradient descent can be quite straightforward, more efficient approaches, *e.g.*, the stochastic approximation approach or a parallel version of the gradient descent [22, 23, 24] can be adopted to improve training efficiency. We will not discuss these algorithms in detail, as it is out of the scope of this paper.

5. Evaluation

In this section, we evaluate our system using a real-world data set. We start by comparing the model performance with three state-of-the-art baselines.

Algorithm 1 PE Solver

Require:

Maximum iteration number S, convergence threshold ϵ;

User feature matrix D, video feature matrix S;

Sparse user-video matrix R.

Parameters set $P = \{p_* | p_* = \{\mu, B_u, B_v, U, V, P, Q\}\}$

Ensure:

Completed user-video matrix $\hat{R}$.

$s \leftarrow 1$;

while $s \leq S$ **and** $L^{(s)} - L^{(s+1)} > \epsilon$ **do**

  $\gamma \leftarrow 1$;

  Compute current residual error by Eq. 7;

  Compute the gradients $\nabla_* L$by Eq. 6;

  while $L(p_* - \gamma \nabla_{p_*} L) \geq L(p_*)$ **do**

    $\gamma = \frac{\gamma}{2}$;

  end while

  $B_u^{(s+1)} = B_u^{(s)} - \gamma \nabla_{B_u}^{(t)}$, $B_v^{(s+1)} = B_v^{(s)} - \gamma \nabla_{B_v}^{(t)}$

  $U^{(s+1)} = U^{(s)} - \gamma \nabla_U^{(t)}$, $V^{(s+1)} = V^{(s)} - \gamma \nabla_V^{(t)}$

  $P^{(s+1)} = P^{(s)} - \gamma \nabla_P^{(t)}$, $Q^{(s+1)} = Q^{(s)} - \gamma \nabla_Q^{(t)}$

  $s = s + 1$

end while

Predict with: $\hat{R} = U \cdot V_s^T + \mu J_{mn} + B_u \cdot J_n^T + J_m \cdot B_v^T$

return $\hat{R}$

Then, we investigate our system by studying how the different parameter settings affect our system's performance.

To conduct the experiments, we have implemented our model in Python 2.7 and ran it on an enterprise server machine with 24 Intel Xeon E5-2420@1.90GHz CPUs, 120 GB memory and 100 TB hard disk.

5.1. Model Performance

To evaluate the effectiveness of personalized models, we employ three widely-used and high-performance machine learning models as baselines: (1) *decision tree regressor*, (2) *random forest* and (3)*gradient boosted regression trees*.

The decision tree is one of the most commonly-used machine learning models in data-driven user engagement analysis [11, 13, 12], while the random forest and gradient boosted regression trees are the ensemble learning models based on decision tree, which introduce bagging and boosting techniques to achieve a better performance [15, 16].

For baseline models, we transform our data set into a flat-table format, (*user features*, *video features*, *download ratio*), in which *download ratio* is used as the prediction label. Then, we find the best parameters for each model via a 10-fold cross validation. The prediction performance is evaluated in terms of *mean absolute error* (MAE) and *root-mean-square error* (RMSE). Given n tuples, let r_i and $\hat{r_i}$ be the real value and predicted value for the i-th tuple, the MAE and RMSE are defined as follows:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|r_i - \hat{r_i}|, \tag{10}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(r_i - \hat{r_i})^2}. \tag{11}$$

Although both MAE and RMSE are metrics for measuring error rate, there are some subtle differences between them. As their names imply, the MAE is a linear metric which means that all the individual errors are weighted equally in the average, while the RMSE gives a relatively high weight to large errors and

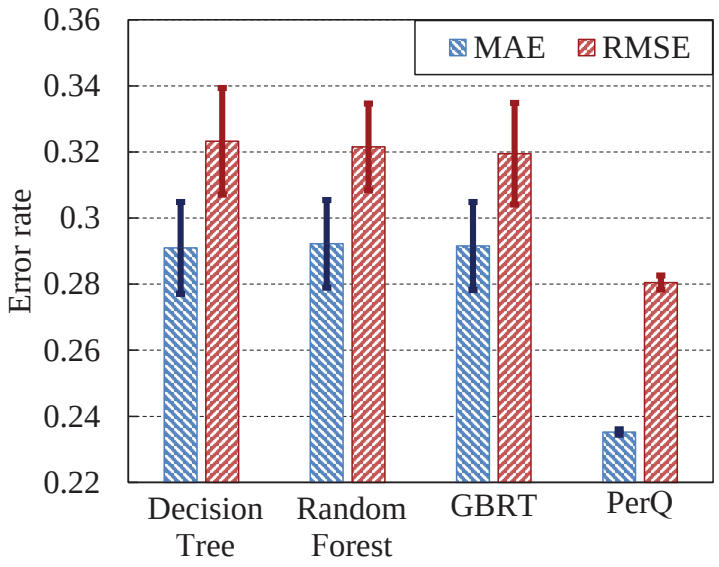

Figure 5: Model performance comparison.

thus it is more useful when large deviations are particularly undesirable. The MAE and RMSE are used together to measure the variance in the individual errors in the prediction. The greater the difference between them, the larger the variance is [19].

Figure 5 shows the comparison of prediction performance of PE and baselines. We observe that, as the *gradient boosted regression trees* and *random forest* leverage the ensemble power of a set of weak learners to build a better model, they slightly outperform the basic *decision tree* model. However, as these baseline models treat all users as a uniform group, they neglect the diversity of user behavior. On the contrary, our model learns individual model for each user with the help of information from user-video interactions, and rich side information from other data sources. When compared with the best of baseline models (*i.e.*, gradient boosted regression trees), the performance improvement is 19.14% and 12.20% in terms of MAE and RMSE, respectively, .

In the following subsections, we will study the performance of our system under different parameter settings.

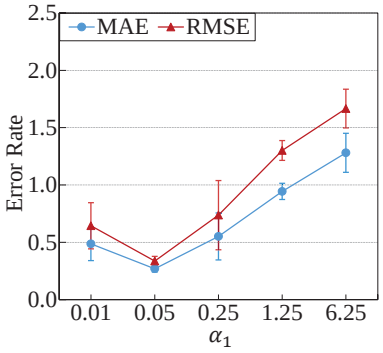

Figure 6: Impact of reconstruction weight α_1 on model performance. The larger α_1 is, the more information is learned from user-feature matrix D, and less emphasis is putted on the approximation of user-video matrix R

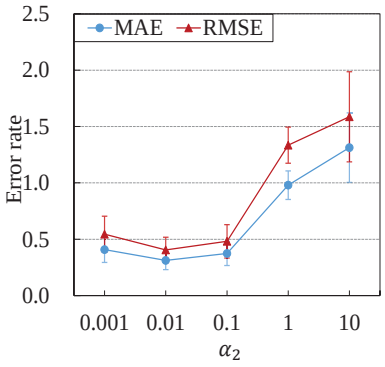

Figure 7: Impact of reconstruction weight α_2. Each value is computed via 10-fold cross validation while fixing α_1 to 0.05

5.2. Impact of Reconstruction Weights

The reconstruction weights, α_1 and α_2, control the importance of corresponding matrix approximation in the loss function and the degree of information propagation. For example, a large α_1 not only implies that more emphasis is placed on the approximation of user feature matrix D in loss function, but also suggests that more information should be learned from D.

To study the impact of these reconstruction weights, we vary the value of one reconstruction weight each time and plot the dynamics of the model performance in terms of the average MAE/RMSE and the standard deviation obtained from the 10-fold cross validation.

Figure 6 demonstrates how the system performance changes as we vary the value of α_1 while fixing $\alpha_2 = 0.01$. We notice that as the value of α_1 increases, the error rate first decreases and then starts to rise. The reason is that, when α_1 is small, our model cannot fully exploit the user-side information to understand the similarity between users and therefore degrades the performance. However, if the value of α_1 is too large, the contribution of the user feature matrix D

would dominate in the loss function. This would restrain the approximation of the user-video matrix, which will eventually downgrade the model performance. The best value of α_1 in our experiment is 0.05.

A similar pattern can also be observed in the analysis of α_2. As shown in Figure 7, the error rate starts to decrease as we enlarge the value of α_2, and more information is propagated from the video feature matrix S. However, if α_2 keeps growing, more errors would be introduced as the approximation of the video feature matrix S dominates in the loss function and thus reduces the importance of filling the missing value in R. The optimal value of α_2 in our experiment is 0.01.

5.3. Impact of Video Clustering Granularity

As we aim to understand the impact of network statistics on user engagement in mobile video services, we utilize the network quality statistics during a video session as features of this video. However, the fluctuation in network quality may lead to a dramatic data explosion and therefore introduce a negative effect to the overall performance. To alleviate this problem, we aggregate video-feature tuples into video templates via agglomerative clustering. The extent of this aggregation is controlled by a clustering granularity c. To understand the impact of video clustering granularity on our system, we validate our system under different cluster granularity settings and present the results in Figure 8.

A small clustering granularity c aggregates more video-feature tuples into a single video template. This can significantly reduce the data size, but also introduce large deviations inside a video template. As a consequence, there will be large prediction errors for video-feature tuples which are *far away* from the video templates. On the other hand, if the value of c is too large, the video-feature tuples are barely aggregated, which again exposes the problem of network quality variations. According to Figure 8, we set $c = 0.5$ for our data set.

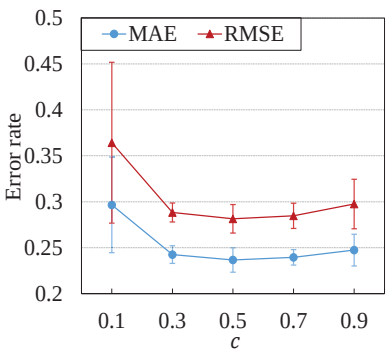

Figure 8: Impact of video clustering granularity c. A larger c enables more video-feature tuples to be aggregated into a single video template and reduce the data sparsity, but also introduce large deviations inside single video template.

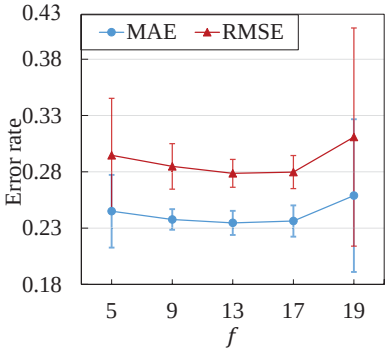

Figure 9: Impact of latent factors f. A small f can significantly compress information into a compact latent space with information loss, while a large f can better capture the underlying pattern in the data, but may aggravate the data sparsity problem.

5.4. Impact of Dimensionality of Latent Factor Space

As our model factorizes each user and video into two vectors in a latent factor space of dimensionality f, a small f can significantly compress information into a compact latent space, which can help conquer the data sparsity problem. However, such information compression may also suffer from information loss and thus cannot achieve the optimal performance. On the other hand, a large f can help better capture the underlying pattern in the data, but bring about a more serious data sparsity problem, which would introduce many errors for users with a small number of watching records.

We plot the average and standard deviation of our system under different values of f in Figure 9. We can observe that, the error rate first decreases when the value of f increases, and achieves an optimum when $f = 13$. If we continue to enlarge the value of f, the data sparsity problem will lead to performance fluctuation.

6. Related Work

Video quality assessment has long been studied in academia. Early works on this area mainly focus on objective QoS metrics, *e.g.*, video encoding rate [2, 25], bitrate [26, 27] or network bandwidth [28, 29], and try to improve user's experience by better QoS provision. However, as the video service is highly user-centric, the practical improvement brought by these works is hard to be validated [10]. To evaluate video quality from the user perspective, many researchers have started to evaluate video quality-of-experience via subjective tests in a controlled environment [6, 7, 10]. The high cost and human participation in subjective tests is inevitable for such works and thus limits the scale of their experiments.

In recent years, the concept of Quality-of-Experience has evolved to the Quality-of-Engagement. The data-driven user engagement analysis for video services has been boosted by the availability of massive data traces from service providers and the fast development of big data processing techniques. Recent literature on data-driven user engagement analysis mainly focus on understanding the influence of different factors on user engagement. In these works, user engagement is quantified from the different perspectives. For example, content providers can quantify user engagement via the viewing time ratio [11], while network service provider may employ the video download ratio [15] as a metric. These metrics also conform with the business models of subscription-based or advertisement-based video services, which is very important from the perspective of service providers. In [11], the authors studied the impact of start-up delay, rebuffer time and encoding bitrate on user engagement. As an extension, in [12, 13], the authors further investigated the impact of types of video, device and connectivity on user engagement and proposed a decision tree-based prediction model to characterize the complicated relationship between user engagement and confounding factors. In [15], Shafig and *et al.* studied how cellular network metrics affect the video download ratio, and predict the download ratio with a regression tree model. However, these existing user engagement models

only quantify the *average* engagement of all users, while user diversity in the engagement pattern has been overlooked. Our work on the personalized user engagement model fills this gap.

7. Conclusion

Accurate and reliable user engagement assessment is the key to optimize the video service quality. Our experiments on real-world data set reveal a significant diversity in user engagement and a uniform user engagement model is not sufficient to characterize the distinctive engagement patterns of different users. To deal with this problem, we propose PE, a personalized user engagement model for mobile videos from the perspective of cellular network providers, in which the user diversity is well addressed. The evaluation results on a real-world data set show that our personalized user engagement model outperforms uniform models with a 19.14% performance gain.

References

[1] Cisco, Cisco visual networking index: Global mobile data traffic forecast update, White Paper.

[2] G. J. Sullivan, T. Wiegand, Rate-distortion optimization for video compression, Signal Processing Magazine, IEEE.

[3] M. Wu, *et al.*, Dynamic resource allocation via video content and short-term traffic statistics, Multimedia, IEEE Transactions on 3 (2) (2001) 186–199.

[4] J. Shin, J. Kim, C. J. Kuo, Quality-of-service mapping mechanism for packet video in differentiated services network, Multimedia, IEEE Transactions on 3 (2) (2001) 219–231.

[5] Q. Zhang, W. Zhu, Y.-Q. Zhang, End-to-end qos for video delivery over wireless internet, Proceedings of the IEEE 93 (1) (2005) 123–134.

[6] ITUT, P. 800: Methods for subjective determination of transmission quality, International Telecommunication Union, Geneva.

[7] ITU-T, Subjective video quality assessment methods for multimedia applications.

[8] H. R. Wu, K. R. Rao, Digital video image quality and perceptual coding, CRC press, 2005.

[9] S. Tao, *et al.*, Real-time monitoring of video quality in ip networks, TON.

[10] Y. Chen, K. Wu, Q. Zhang, From QoS to QoE: A Tutorial on Video Quality Assessment, IEEE Communications Surveys and Tutorials.

[11] F. Dobrian, *et al.*, Understanding the impact of video quality on user engagement, in: SIGCOMM, 2011.

[12] A. Balachandran, *et al.*, Developing a predictive model of quality of experience for internet video, in: SIGCOMM, 2013.

[13] A. Balachandran, *et al.*, A quest for an internet video quality-of-experience metric, in: HotNets, 2012.

[14] S. S. Krishnan, R. K. Sitaraman, Video stream quality impacts viewer behavior: Inferring causality using quasi-experimental designs, in: IMC, 2012.

[15] M. Z. Shafiq, *et al.*, Understanding the impact of network dynamics on mobile video user engagement, in: SIGMETRICS, 2014.

[16] J. H. Friedman, Greedy function approximation: a gradient boosting machine, Annals of statistics.

[17] China unicom.
URL `https://en.wikipedia.org/wiki/China_Unicom`

[18] A. Singh, G. J. Gordon, Relational learning via collective matrix factorization, in: SIGKDD, 2008.

[19] C. Bishop, Pattern recognition and machine learning, springer, 2006.

[20] L. Kaufman, P. J. Rousseeuw, Finding groups in data. an introduction to cluster analysis, Applied Probability and Statistics 1.

[21] Y. Koren, R. Bell, Advances in collaborative filtering, in: Recommender systems handbook, 2011.

[22] T. Zhang, Solving large scale linear prediction problems using stochastic gradient descent algorithms, in: ICML, 2004.

[23] M. Zinkevich, *et al.*, Parallelized stochastic gradient descent, in: NIPS, 2010.

[24] B. Recht, *et al.*, Hogwild: A lock-free approach to parallelizing stochastic gradient descent, in: NIPS, 2011.

[25] Z. Chen, K. N. Ngan, Recent advances in rate control for video coding, Signal Processing: Image Communication.

[26] G. Sullivan, T. Wiegand, Rate-distortion optimization for video compression, Signal Processing Magazine, IEEE.

[27] Z. Chen, K. N. Ngan, Recent advances in rate control for video coding, Image Commun. 22.

[28] S. Chong, S. qi Li, J. Ghosh, Predictive dynamic bandwidth allocation for efficient transport of real-time vbr video over atm, JSAC.

[29] M. Wu, *et al.*, Dynamic resource allocation via video content and short-term traffic statistics, TMM.