Individual Preference Probability Modeling for Video Content in Wireless Caching Networks

Ming-Chun Lee and Andreas F. Molisch
Ming Hsieh Department of Electrical Engineering
University of Southern California
Los Angeles, CA, USA
mingchul@usc.edu, molisch@usc.edu

Nishanth Sastry and Aravindh Raman
Department of Informatics
King's College London
London, UK
nishanth.sastry@kcl.ac.uk, aravindh.raman@kcl.ac.uk

***Abstract*—Caching of video files at the wireless edge, i.e., at the base stations or on user devices, is a key method for improving wireless video delivery. While *global* popularity distributions of video content have been investigated in the past, and used in a variety of caching algorithms, this paper investigates the *statistical modeling of the individual user preferences*. With individual preferences being represented by probabilities, we identify their critical features and parameters and propose a novel modeling framework as well as a parameterization of the framework based on an extensive real-world data set. Besides, an implementation recipe for generating practical individual preference probabilities is proposed. By comparing with the underlying real data, we show that the proposed models and generation approach can effectively characterize individual preferences of users for video content.**

I. Introduction

Data traffic generated by the demand for video content in wireless networks has approximately doubled every year and is expected to continue to grow in the next several years [1], [2]. Conventional approaches, such as using more efficient transceivers, densifying infrastructure, and/or using more spectrum, for supporting the increasing traffic are deemed insufficient or too expensive [2], [3]. An important alternative that has emerged in the past years is video caching at the wireless edge. Leveraging unique features of video popularity and the low cost of storage resources, video caching has shown its potential and drawn wide attention [1]–[4].

Video content caching has been discussed in different networks with different equipments being used as the storage resources [1]–[4]. Femtocaching and base station (BS) caching use storage resources in helper nodes and BSs to cache video content and provide the ability to immediately serve users without using backhaul [5]–[7]. Video content cached directly in mobile devices provides a more direct and a higher density caching apporach [8]–[10]. As the caching video content is in mobile devices, mobile users can either directly reach the video content from their own storage without consuming any resource [8] or exploit device-to-device communications to access video content with low cost [8]–[10]. The combination of storage on user devices together with coded multicast has also been widely explored [11].

Although designs for improving wireless video content caching have been widely explored, most of the literature adopts a homogeneous popularity model, i.e., assumes all users have the same file popularity distribution for deciding the desired video content [12]. Clearly this assumption violates the intuition that different users have different tastes and preferences. Therefore these designs are restricted to some extent due to lack of considering individual user preference. Note that modeling the individual preferences of a particular user, also known as the "Netflix challenge", has been investigated intensely [16], [17]. However, this is different from the need to find *statistics* of individual user distributions.

Approaches exploiting individual preference for caching or delivering content have just recently been discussed [12]–[16]. By exploiting individual preference, designs of wireless caching networks can be refined and improved [12]–[14]. Besides, analyses with individual preference can offer fundamental insights that might further enhance the system or strategy designs [15], [16]. However, to the best of our knowledge, there does not exist any statistical model for the individual user distributions based on real-world data. The current paper aims to fill this gap.

Our model uses hierarchies of probabilities to represent preferences of users. Empirically, video files can be categorized into genres according to their features, and users might have strong preferences toward a few genres [16]. The overall request probability of a user for a file is then modeled as the probability that a user wants a specific genre, and the popularity of a file within this genre. Since the individual preference probabilities of users can be described by the individual popularity distributions and ranking orders, statistics of them are respectvely investigated using the genre-based structure. We note that, in this paper, we implicitly denote the distribution as the rank-frequency distribution when we use the term: *popularity distribution*. We will extract the models and parameterization for these different statistics.

Such modeling and parameterization has to be based on real-world data to be meaningful. We are thus using data from an extensive dataset collected in the U.K. in 2014, namely the usage of the BBC iplayer [15], [16]. By observing the real data, we identify several important aspects of characterizing individual preferences, and propose the modeling framework for individual preference probabilities. By following the modeling framework, an individual preference probability generation approach is also proposed. We validate the proposed modeling and generation approach with real data. The validation results demonstrate that proposed modeling and generation approach can effectively reproduce important features and statistcs of the individual preference. Therefore it can serve for designing, optimizing, analyzing, modeling, and simulating wireless caching networks.

The remaining paper is organized as follows. Section II introduces the basic modeling concepts and describes the necessary tools for manipulating the dataset. Main modeling results are provided in Sections III and IV. We propose the individual preference probability generation approach in Section V. Section VI presents the conclusions.

II. Individual Preference Probability Modeling and Dataset Preparations

A. Modeling on Individual Preference Probability

In this work, we consider the individual user probability distributions, which are defined as the probability that a specific user will in the future request a specific file for watching; multiple views by the same user are thus ignored (i.e., treated the same as single viewing). Since different users could have different preferences, preference probabilities of different users for the same file could be different. In this work we consider each file can be categorized into a genre, and there are G genres in the library. Therefore denoting M_g as the number of files in genre g, the total number of files in the library is given by $\sum_{g=1}^{G} M_g$. Given this library, we denote the preference probability of the file m in genre g for user k as $p_{g,m}^k$. Then the following properties must hold[1]: $0 \leq p_{g,m}^k \leq 1, \forall g, m, k$ and $\sum_{g=1}^{G} \sum_{m=1}^{M_g} p_{g,m}^k = 1, \forall k$.

To characterize individual preference probabilities of users, two important features need to be characterized: individual popularity distrbutions of files and individual ranking orders of files. Different individual popularity distrbutions represent different *concentration* rates of popularity distributions that different users might have, and different individual ranking orders represent different preferences for files by ranking files differently. Here we provide a simple example for elaboration. We consider two users with different preferences. Suppose that $G = 1$ and $M_1 = 3$. Therefore there are three files in the library. Then assume we know $p_{1,1}^1 = 0.5$, $p_{1,2}^1 = 0.3$, $p_{1,3}^1 = 0.2$; and $p_{1,1}^2 = 0.05$, $p_{1,2}^2 = 0.7$, $p_{1,3}^2 = 0.25$. Note that these six popularity values are a complete description, but obviously such a description becomes impossible to handle when considering thousands of files and millions of users. It can be observed that their popularity distributions are totally different. Besides, the ranking orders are different, namely 1, 2, 3 and 2, 3, 1, repsectively. By using the example, it can be observed that the differences between preferences of users can be fully described by the differences of individual popularity disributions and individual ranking orders.

[1]We note that, by using this model, we implicitly consider every user being equally important. However, from system's point of view, different weighting on different users according to certain strategy might be desired. Investigations of such weighted models from system's point of view are also important and are considered as a future direction.

To avoid confusions, in the following sections, we use global popularity/probability of genres/files to denote the popularity/probability of genres/files computed by taking all users into consideration. As the counterpart, the individual popularity/probability of genres/files is used to denote the popularity/probability computed by considering only a single specific user. In addition, without loss of generality, we consider the indices of genres to follow the decending order of the global popularities of genres, i.e., the global popularity of genre g is larger than the global popularity of genre $g+1$ for all $1 \leq g \leq G$.

B. Dataset Descriptions and Preprocessing

This work uses an extensive set of real-world data, namely the dataset of the BBC iplayer [15], [16]. The BBC iPlayer is a video streaming service provided by BBC (British Broadcasting Corporation). Video and radio content are provided for a number of BBC channels without charge. Content on the iPlayer is basically available for up to 30 days depending on the policies. We consider the month-long dataset accommodating up to 192,120,311 recorded access sessions of June, 2014. In each record, access information for the video content contains two important column: *user id* and *content id*. *user id* is based on the long-term cookies that unqiuely (in an anonymized way) identify users. *content id* is the specific identity that uniquely identifies each video content separately. Although there are certain exceptions, *user id* and *content id* can generally help us identify the user and the video content of each access. In addition to access identifications, video files in the BBC iplayer are annotated with one or more genres. The annotated genres for each video content are used to help us identify users' preferences. Notice that there are certain files that are not annotated with any genre. We simply filter them out, as described in the following paragraph. Detailed descriptions of the BBC iplayer dataset can be found in [15], [16].

To facilitate the investigation, preprocessing is conducted on the dataset. To be specific, we concentrate our investigation on regular (frequent) users. To define a regular user, we first define the unique access. By observations, we notice that a user could access the same file multiple times, possibly due to temporary disconnnections from Internet and/or due to temporary pauses raised by users when moving between locations. Since a user is generally unlikely to access the same video after finishing to watch the video within the period of a month, we consider multiple accesses made by the same user to the same file as a single unique access. A regular user is a user with more than 100 unique accesses in a month. An analysis that includes less frequent users will be presented in [19].

As described previously, a file could be annotated with one or several genres. The genre-wise classification is the foundation for characterizing preferences of users in our work. Hence if a file could not be classified into any genre, i.e., if no genre is annotatated on the file, the file is filtered out during the preprocessing.