Probabilistic Caching Policy for Categorized Contents and Consecutive User Demands

Minseok Choi[†*], Dongjae Kim[†], Dong-Jun Han[†], Joongheon Kim[*], and Jaekyun Moon[†]
[†]School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea
[*]School of Computer Science and Engineering, Chung-Ang University, Seoul, Korea
E-mails: {ejaqmf,codong,djhan93}@kaist.ac.kr, joongheon@cau.ac.kr, jmoon@kaist.edu

***Abstract*—In wireless caching networks, each user generally consumes more than one content in a row, and the number of consecutive demands could vary for different users. In addition, popular contents are usually classified into several categories. In this case for consecutive user demands, the content popularity model largely depends on the previously consumed contents, i.e., contents that belong to the same category as the previously consumed content would be highly popular. Based on this observation, this paper proposes an optimal probabilistic caching policy for consecutive user demands in categorized contents. The proposed caching scheme maximizes the minimum of the success probabilities for content delivery of all users when individual users request different numbers of contents in a row. Comparing with the content placement optimized for one-shot request, intensive numerical results verify the impacts of categorized contents and consecutive user demands on the caching policy.**

I. Introduction

Tens of exabytes of global data traffic are being handled now on a daily basis [1]. In many mobile services, a relatively small number of popular contents is requested at ultra high rates, e.g., on-demand streaming services [2], [3]. In this respect, most of user demands for diverse multimedia contents are overlapped and repeated. To deal with this issue, wireless caching technologies have been studied, wherein the base station (BS) or the server pushes popular contents for off-peak time to cache-enabled nodes so that these nodes provide popular contents directly to nearby mobile users [4]–[6]. Since the wireless caching allows popular contents closer to users, it also has an advantage of reducing the content delivery delay.

To take full advantage of wireless caching, many caching helpers, which act as small BSs, should be installed near users [4]. Furthermore, a device-to-device (D2D)-assisted caching network has been studied in [7]–[9], where mobile devices can store popular contents and directly respond to the file requests of neighboring users [10]. In practice, both caching helpers and cache-enabled devices have finite storage size owing to cost issues. Therefore, the system should determine which content is better to be stored in helpers or cache-enabled devices. This problem is commonly known as the content placement problem.

The goal of the content placement problem is to find optimal caching policies according to the popularity distribution of contents and network topology. With a fixed network topology, the caching schemes to minimize the average downloading delay and the average bit error rate have been proposed in [11] and [12], respectively. Considering channel fading effects, the deterministic caching schemes are proposed to minimize the average delay [13] and to minimize the outage probability in relay networks [14]. In addition, the authors of [15] proposed a joint caching and routing technique to minimize the requests routed to the BS, while [16] discussed cooperative caching and delivery for minimizing average downloading latency.

In stochastic wireless caching networks, there exist several research efforts on probabilistic content placement introduced in [17]. Many probabilistic caching methods have been proposed depending on various optimization goals, e.g., maximization of cache hit probability [17], cache-aided throughput [18], average success probability of content delivery [19], average successfully enjoyable content quality [20].

However, these previous research results on the content placement problem do not consider the consecutive user demands for categorized contents. The content popularity for one-shot request is commonly modeled by the Zipf distribution [7]. However, if the user requests multiple contents continuously and contents are categorized, the popularity model will not be likely to follow the Zipf distribution. For example, when a user begins to watch a video, the popularity on the first content is random, e.g., Zipf distribution. However, the popularity of the next video largely depends on the previously watched video and its category when the user chooses the video right after watching the previous one. We can expect that the popularity of contents whose category is the same as that of previously consumed contents will be high. In this respect, this paper proposes a probabilistic caching policy for consecutive content demands in wireless caching helper networks.

The main contributions are as follows:

- Different from most results on the content placement problem in which only one-shot request is considered, consecutive user demands and different numbers of content requests for users are considered. In practice, even if users enjoy the same service, there are heavy users who consume a lot of contents in a row and users who do not.
- The concept of a content category and the different popularity model for consecutive user demands for categorized contents are captured in the proposed system model. Considering consecutive user demands, the popularity model for the categorized contents depends on the previously consumed contents.
- The iterative algorithm for finding the optimal probabilistic caching policy for categorized contents and consecutive user demands is proposed. The proposed caching

scheme maximizes the minimum of the successful delivery rates of all users. The proposed iterative algorithm can guarantee to make the optimal solution converge.
- Numerical results show the impacts of categorized contents and consecutive user demands on the caching policy. Performance gains of the proposed scheme increase as contents in the same category become highly relevant and the number of consecutive requested contents grows.

The rest of the paper is organized as follows. The system model is described in Section II, and the average successful delivery rate for consecutive user demands is derived in a mathematical form in Section III. In Section IV, the optimal probabilistic caching policy is proposed. The numerical results are shown in Section V and Section VI concludes the paper.

II. System Model

This section describes the wireless caching network and the content popularity model. We consider the scenario in which users consume multiple contents in sequence and contents are grouped into several categories.

A. Wireless Caching Network

This paper considers a cellular model where multiple caching helpers exist and users request a particular cached content from a library $\mathcal{F}$. Suppose that a library $\mathcal{F}$ consists of F contents and all contents are with normalized unit sizes. For the contents of different sizes, each content can be partitioned into small chunks of the same size and each chunk can be considered as an individual content. Users search through the helper candidates that cache the requested content within a radius of R, as shown in Fig. 1, and each user selects one of the candidates for content delivery.

The caching helpers have the finite storage size of M, which means only M contents can be cached in each helper. In practice, $F > M$, therefore caching helpers cannot store all of contents in $\mathcal{F}$. If there is no helper caching the requested content within the radius R from the user, the server which has the whole library can deliver the desired content via a cellular link. Since the caching helpers are usually much closer to the content-requesting users than the server, the users are assumed to prefer downloading the content from the caching helpers rather than directly from the server, due to transmission delay. Therefore, direct transmission from the server is not considered in this paper.

Different from most of existing works on wireless caching policy, this paper allows each user to request multiple contents consecutively. For example, on-demand streaming users usually start watching videos with the category they clearly want to see, i.e., sports highlights. They would watch multiple videos in the similar category in a row, and each user tends to consume different numbers of videos. Therefore, this paper considers L types of users and a type-l user requests l cached contents in a row from nearby helpers for $l = 1, \cdots, L$. For example, in Fig. 1, the type-L user is scheduled to exploit the wireless caching network, and the user can receive the desired contents from several nearby helpers.

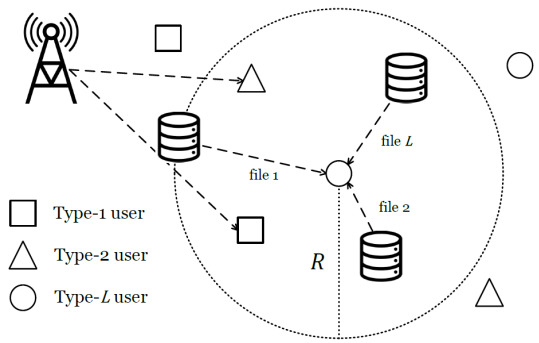

Fig. 1. Wireless caching network model

The caching helpers and users are modeled using the independent Poisson point processes (PPPs) with intensity λ and λ_u, respectively. Assume that each type of users is distributed with the same intensity. This paper utilizes the probabilistic caching placement method [7] for caching helpers to cache file i with probability p_i. Suppose that the system does not allow any additional link activation within the radius R of the user who is already downloading the content from certain helper. By taking R sufficiently large and/or exploiting orthogonal resources for each user's coverage, the system can guarantee negligible interference among multiple links.

The Rayleigh fading channel is assumed for the wireless links from users to their caching helpers. Denote the channel with $h = \sqrt{D}g$ where $D = 1/d^{\alpha}$ controls path loss with d being the user-device distance and α being the path loss exponent. In addition, g represents a fast fading component with complex Gaussian distribution $g \sim CN(0, 1)$.

B. Content Popularity Model

Each content $i \in \mathcal{F}$ has a popularity probability f_i, which follows the Zipf distribution [17]: $f_i = i^{-\gamma} / \sum_{j=1}^{F} j^{-\gamma}$ where γ denotes the popularity distribution skewness. Since this paper considers users who request multiple contents in a row, the popularity model for consecutive content requests is required. If a user is consuming a certain content, then this user will be highly likely to request another content which belongs to the same category as that of the previous content. On the other hand, popularities of contents in other categories would be very low. This popularity model depends on the previous content and it is clearly different from the popularity model for the one-shot content request, i.e., Zipf distribution. Thus, when a user is consuming content i, the popularity of the next content j is denoted by $f_{j|i}$ and $f_{j|i} \neq f_j$.

Let a library $\mathcal{F}$ is divided into K categories, denoted by $\mathcal{G}_k$ for $k = 1, \cdots, K$ and each category consists of N contents. $\mathcal{G}(i)$ denotes the index of the category which content i belongs to. Since the user is more likely to request the next content whose category is the same as the current content than other categories, $f_{j|i} > f_{k|i}$ for all j and k satisfying $\mathcal{G}(j) = \mathcal{G}(i)$ and $\mathcal{G}(k) \neq \mathcal{G}(i)$. In addition, the popularity model for contents in each category could be independently designed, if there is no correlation among categories. The following

sections can be applied to any popularity model of $f_{j|i}$, e.g., uniform distribution or Mandelbrot-Zipf distribution [21].

III. Average Successful Delivery Rate for Consecutive User Demands

In this section, the successful delivery rate for consecutive user demands is mathematically derived, and the problem for finding the caching policy is presented.

A. The Average Successful Delivery Rate of Type-L User

Note that the type-L user requests L contents in a row. Therefore, the outage event occurs for the type-L user if at least one of L content requests fails to be delivered. According to Slivnyaks theorem, we consider a typical user located at the origin and the statistics of the typical user represent those of any other user generated by a PPP with the same intensity. If the user desires content i and there are multiple helper candidates caching content i, it is reasonable for the user to download the content from the helper whose channel condition is the strongest among the candidates.

Let R_i be the data rate of the typical type-L user to receive the desired content i from the helper whose channel is the strongest among helper candidates caching the content i. Denoting the Rayleigh fading channel from the user to the nearest caching helper for downloading content i by h_i, the data rate of the user for downloading file i is given by

$$R_i = \mathcal{B}\log_2\Big(1+\frac{|h_i|^2}{\sigma^2}\Big), \tag{1}$$

where $\mathcal{B}$ is the bandwidth, assuming a unit transmit power and a normalized noise variance of σ^2.

Then, the average success probability of type-L user for content delivery is given by

$$P_L^o = \sum_{i_1}\cdots\sum_{i_L} f_{i_1} f_{i_2|i_1}\cdots f_{i_L|i_{L-1}} \cdot \prod_{i=i_1}^{i_L} \Pr\{R_i \geq \rho\}, \tag{2}$$

where i_l represents the index of the l-th content requested by the user and ρ is the threshold for data rates. Although it is omitted in (2), $i_l \neq i_1,\cdots,i_{l-1}$ is assumed for all $l=2,\cdots,L$ as mentioned in (23). In other words, we suppose that the previous consumed content will not be requested again in future because most of devices usually store contents consumes in recent in their local cache memory.

Since the channel power $|h_i|^2$ follows the chi-squared distribution, i.e., Nakagami-1 fading channel and is the strongest among those of caching helper candidates, according to Lemma 1 in [19], the reliable transmission probability can be obtained by

$$P\{R_i \geq \rho\} = 1-\exp\Big\{-\kappa p_i\Big(\frac{1}{\sigma^2(2^{\rho/\mathcal{B}}-1)}\Big)^{\frac{2}{\alpha}}\Big\}, \tag{3}$$

where $\kappa = \pi\lambda\Gamma(\frac{2}{\alpha}+1)$. Thus, the average successful delivery rate of type-L user can be written by

$$P_L^o = \sum_{i_1}\cdots\sum_{i_l} f_{i_1} f_{i_2|i_1}\cdots f_{i_L|i_{L-1}} \cdot \prod_{i=i_1}^{i_L}(1-e^{-Cp_i}), \tag{4}$$

where $C = \kappa\Big(\frac{1}{\sigma^2(2^{\rho/\mathcal{B}}-1)}\Big)^{\frac{2}{\alpha}}$.

B. Problem Formulation

The goal of this paper is to find the caching probabilities which maximize the average successful delivery rates of all types of users with the maximum number L of consecutive content requests. Then, the optimization problem is formulated to maximize the minimum average success probability among all types of users as follows:

$$\mathbf{p}^* = \underset{p_i, i=1,\cdots,F}{\arg\max}\Big[\min\{P_1^o,\cdots,P_L^o\}\Big] \tag{5}$$

$$\text{s.t.} \quad \sum_{i=1}^{N} p_i \leq M \tag{6}$$

$$0 \leq p_i \leq 1. \tag{7}$$

where (6) results from the finite memory size of caching helpers based on the probabilistic caching method in [17].

IV. Optimal Probabilistic Caching Policy

This section provides key lemmas and finds the probabilistic caching policy by solving the problem of (5)-(7).

A. Key Lemmas and Problem Re-Organization

The optimization problem of (5)-(7) can be re-organized via the following lemmas. By using Lemma 1, the max-min problem of (5)-(7) can be transformed into a simple convex maximization problem. In addition, the Lemma 2 turns the inequality constraint (6) into the equality constraint.

Lemma 1. $P_l^o > P_m^o$ for any $l, m \in \{1,\cdots,L\}$ and $l < m$.

Proof: By showing $P_l^o > P_{l+1}^o$, this can be proved.

$$\begin{aligned}
&P_l^o - P_{l+1}^o \\
&= \sum_{i_1}\cdots\sum_{i_l} f_{i_1} f_{i_2|i_1}\cdots f_{i_l|i_{l-1}} \cdot \Bigg[\prod_{i=i_1}^{i_l}(1-e^{-C_i p_i}) \\
&\quad - \sum_{i_{l+1}} f_{i_{l+1}|i_l}\Bigg\{\prod_{i=i_1}^{i_l}(1-e^{-C_i p_i})(1-e^{-C_{i_{l+1}} p_{i_{l+1}}})\Bigg\}\Bigg] \\
&= \sum_{i_1}\cdots\sum_{i_l} f_{i_1} f_{i_2|i_1}\cdots f_{i_l|i_{l-1}} \\
&\quad \times \Bigg[\prod_{i=i_1}^{i_l}(1-e^{-C_i p_i})\Big(1-\sum_{i_{l+1}} f_{i_{l+1}|i_l}(1-e^{-C_{i_{l+1}} p_{i_{l+1}}})\Big)\Bigg] \\
&> 0.
\end{aligned}$$

Since $\sum_{i_{l+1}} f_{i_{l+1}|i_l} = 1$, the second equality and the last inequality are satisfied. ∎

Lemma 2. The optimum vector $\mathbf{p}^* = (p_1^*,\cdots,p_F^*)^T$ satisfies

$$\sum_{i=1}^{F} p_i^* = M. \tag{8}$$

Proof: Assume $\sum_{i=1}^{F} p_i^* < M$, then $\exists\epsilon > 0$ such that $\sum_{i=1}^{F} p_i^* + \epsilon \leq M$ and $p_k^* + \epsilon \leq 1$ for certain $k \in \{1,\cdots,F\}$. Let $\mathbf{p}' \triangleq (p_1^*,\cdots,p_k^*+\epsilon,\cdots,p_F^*)^T$, then since P_L^o is an increasing function of any p_i for $i=\{1,\cdots,L\}$,

$$P_L^o(\mathbf{p}') < P_L^o(\mathbf{p}^*).$$

Thus, it obviously leads to contradiction. ■

According to Lemmas 1 and 2, the max-min optimization problem of (5)-(7) can be transformed into the following maximization problem:

$$\mathbf{p}^* = \underset{p_i, i=1,\cdots,F}{\arg\max}\ P_L^o \tag{9}$$

$$\text{s.t.} \quad \sum\nolimits_{i=1}^{F} p_i = M \tag{10}$$

$$0 \le p_i \le 1 \tag{11}$$

and this optimization is convex since P_L^o is convex by (4).

B. Subproblem for Optimization of Two Contents and Iterative Algorithm

Since P_L^o is a multivariable function and consists of many exponential terms, the iterative algorithm is used to find optimal caching probabilities. The subproblem with respect to two variables is formulated by considering the other variables as constants, i.e., let p_m and p_n be caching probabilities to be optimized and make the other probabilities $\{p_i\}_{i\neq m,n}$ be fixed. In terms of p_m and p_n, P_L^o of (4) can be divided into four different parts as follows:

$$\begin{aligned} P_L^o = {} & a_{m,n}(1-e^{-Cp_m})(1-e^{-Cp_n}) + \\ & b_{m,n}(1-e^{-Cp_m}) + d_{m,n}(1-e^{-Cp_n}) + e_{m,n}, \end{aligned} \tag{12}$$

where $a_{m,n}$, $b_{m,n}$, $d_{m,n}$, and $e_{m,n}$ are constants consisting of system parameters, e.g., f_i, $f_{j|i}$, e^{-Cp_k} for all $i,j,k \in \{1,\cdots,F\}$ and $k \neq i,j$. The first part of $(1-e^{-Cp_m})(1-e^{-Cp_n})$ represents the event where the user requests both contents m and n. Similarly, the second and third parts of $(1-e^{-Cp_m})$ and $(1-e^{-Cp_n})$ correspond to the events where the user requests only one content between m and n. The last constant term $e_{m,n}$ is obtained for when the user does not request contents m and n. Therefore, $b_{m,n}$ is given by

$$\begin{aligned} b_{m,n} = {} & f_m \sum_{i_2} \cdots \sum_{i_L} f_{i_2|m} \cdots f_{i_L|i_{L-1}} \\ & + \sum_{i_1} \sum_{i_3} \cdots \sum_{i_L} f_{i_1} f_{m|i_1} \cdots f_{i_L|i_{L-1}} \\ & + \cdots + \sum_{i_1} \cdots \sum_{i_{L-1}} f_{i_1} \cdots f_{i_{L-1}|i_{L-2}} f_{m|i_{L-1}}. \end{aligned} \tag{13}$$

In addition, $a_{m,n}$, $d_{m,n}$, and $e_{m,n}$ can be obtained in a similar way of the procedure to obtain $b_{m,n}$. Then, the subproblem for finding the optimal p_m and p_n is formulated as follows:

$$\{p_m^*, p_n^*\} = \underset{p_m,p_n}{\arg\min}\ \mathcal{M}_{(p_m,p_n)} \tag{14}$$

$$\text{s.t.} \quad p_m + p_n = q_{m,n} = M - \sum\nolimits_{i=1, i\neq m,n}^{F} p_i \tag{15}$$

$$0 \le p_m, p_n \le 1, \tag{16}$$

where

$$\mathcal{M}_{(p_m,p_n)} = (a_{m,n}+b_{m,n})e^{-C\cdot p_m} + (a_{m,n}+d_{m,n})e^{-C\cdot p_n} \tag{17}$$

and this $\mathcal{M}_{(p_m,p_n)}$ is obtained by removing the constant terms and reversing the sign. Since $\{p_i\}_{i\neq m,n}$ are fixed, $p_m + p_n$ also becomes a constant. The following proposition provides the solution of the above subproblem.

Proposition 1. The optimal solution of the problem (14)-(16) is as follows:

$$\{p_m^*, p_n^*\} = \begin{cases} \{\tilde{p}_m, \tilde{p}_n\} & \text{if } 0 \le \tilde{p}_m, \tilde{p}_n \le 1 \\ \underset{\{p_m,p_n\}}{\arg\min}\{\mathcal{M}_{(0,q_{m,n})}, \mathcal{M}_{(q_{m,n},0)}\} & \text{elseif } q_{m,n} < 1 \\ \underset{\{p_m,p_n\}}{\arg\min}\{\mathcal{M}_{(1,q_{m,n}-1)}, \mathcal{M}_{(q_{m,n}-1,1)}\} & \text{elseif } q_{m,n} \ge 1, \end{cases} \tag{18}$$

where

$$\tilde{p}_m = \frac{1}{2C}\log\frac{b_{m,n}}{d_{m,n}} + \frac{1}{2}q_{m,n} \tag{19}$$

$$\tilde{p}_n = \frac{1}{2C}\log\frac{d_{m,n}}{b_{m,n}} + \frac{1}{2}q_{m,n}. \tag{20}$$

Proof: According to the arithmetic-geometric mean inequality, the lower bound on $\mathcal{M}_{(p_m,p_n)}$ is given by

$$b_{m,n}e^{-C\cdot\tilde{p}_m} + d_{m,n}e^{-C\cdot\tilde{p}_n} \ge 2\sqrt{b_{m,n}d_{m,n}e^{-C\cdot q_{m,n}}}. \tag{21}$$

The equality holds if and only if

$$b_{m,n}e^{-c\cdot\tilde{p}_m} = d_{m,n}e^{-c\cdot\tilde{p}_n}. \tag{22}$$

Since $\tilde{p}_m = q_{m,n} - \tilde{p}_n$, $\tilde{p}_m$ and $\tilde{p}_n$ are found as given by (19) and (20), respectively. For (16), $\tilde{p}_m$ and $\tilde{p}_n$ become the optimal solution only when $0 \le \tilde{p}_m, \tilde{p}_n \le 1$. Otherwise, the four boundary conditions are compared as follows: 1) $p_m = 0$ and $p_n = q_{m,n}$, 2) $p_m = q_{m,n}$ and $p_n = 0$, 3) $p_m = 1$ and $p_n = q_{m,n} - 1$, and 4) $p_m = q_{m,n}$ and $p_n = q_{m,n} - 1$. Thus, the optimal solution (18) is obtained. ■

Then, a multivariable function can be optimized by iteratively optimizing the subset of variables if the convergence is guaranteed. To find the optimal $\mathbf{p}^* = [p_1^*, \cdots, p_F^*]$, the subproblem of (14)-(16) can be iteratively applied for all combinations of m and n, where $m,n \in \{1,\cdots,F\}$ and $m \neq n$. The details are given in Algorithm 1.

If a sequence is nonincreasing and has a lower bound, this sequence converges. We find the minimum of the dual-variable problem of (14)-(16) in each iteration, and the sequence of the updated objective values $\mathcal{M}_{(p_m,p_n)}$ is generated. Since this sequence is non-increasing and the average success probability has a trivial lower bound of 0, i.e., $0 \le P_L^o$, the convergence of the proposed algorithm is guaranteed.

V. Numerical Results

In this section, we numerically show the impacts of categorized contents and consecutive user demands on the caching policy. In addition, we show how caching probabilities and the average success probabilities for content delivery are affected by various network parameters. For simulation settings, $F = 25$ contents are grouped into $K = 5$ categories and each consists of $N = 5$ contents. For simplicity, uniform distribution is assumed for the popularity model of $f_{j|i}$, as

Algorithm 1 Iterative algorithm for the optimization problem of (9)-(11)

Precondition:

- M: memory size
- F: the number of contents

$p_i^* = \frac{M}{F}$ for all $i \in \{1, \cdots, F\}$
for $\forall (m,n) \in \{1, \cdots, F\} \times \{1, \cdots, F\}$ and $m \neq n$ **do**
  $q_{m,n} = M - p_m^* - p_n^*$
  Find $\tilde{p}_m$ and $\tilde{p}_n$ according to (19) and (20).
  if $0 \leq \tilde{p}_m, \tilde{p}_n \leq 1$ **then**
    $p_m^* \leftarrow \tilde{p}_m$ and $p_n^* \leftarrow \tilde{p}_n^*$
  else if $q_{m,n} < 1$ **then**
    $\{p_m^*, p_n^*\} \leftarrow \underset{\{p_m, p_n\}}{\arg\min} \{\mathcal{M}_{(0,q_{m,n})}, \mathcal{M}_{(q_{m,n},0)}\}$
  else if $q_{m,n} \geq 1$ **then**
    $\{p_m^*, p_n^*\} \leftarrow \underset{\{p_m, p_n\}}{\arg\min} \{\mathcal{M}_{(1,q_{m,n}-1)}, \mathcal{M}_{(q_{m,n}-1,1)}\}$
  end if
end for

follows:

$$f_{j|i} = \begin{cases} \frac{q}{(N - \mathcal{N}(\mathcal{G}(i)))} & \text{if } \mathcal{G}(j) = \mathcal{G}(i) \\ \frac{(1-q)}{(F - N - \mathcal{N}(\bigcup_{n \neq i} \mathcal{G}(n)))} & \text{if } \mathcal{G}(j) \neq \mathcal{G}(i) \end{cases}, \quad (23)$$

where q is the probability of requesting the content in the same category of the previous content. In addition, $\mathcal{N}(k)$ denotes the number of contents in $\mathcal{G}_k$ which the user has already consumed before. Suppose that the lower content index indicates the more popular content, i.e., $f_i > f_j$ for $i < j$, and the content lists of K categories are as follows: $\mathcal{G}_1 = \{1, 6, 11, 16, 21\}$, $\mathcal{G}_2 = \{2, 7, 12, 17, 22\}$, $\mathcal{G}_3 = \{3, 8, 13, 18, 23\}$, $\mathcal{G}_4 = \{4, 9, 14, 19, 24\}$, and $\mathcal{G}_5 = \{5, 10, 15, 20, 25\}$. Assume that $M = 3$, $\mathcal{B} = 1MHz$, $\rho = 1Mbps$, and $\alpha = 3$. In addition, $L = 4$, $\lambda = 0.2$, and $q = 0.9$ are used, unless otherwise noted.

For comparison purposes, the probabilistic caching policy optimized for the case of $L = 1$ to maximize the average success probability for content delivery [19] is considered. Fig. 2 shows the caching probabilities of all contents in $\mathcal{F}$. In Fig. 2, caching probabilities obtained by the comparison scheme in [19] depend on the content popularity for one-shot request, i.e., Zipf distribution. On the other hand, caching probabilities of the proposed scheme are largely influenced by the popularity model for consecutive content requests. Caching probabilities of contents in $\mathcal{G}_1$ are the highest among all categories and the contents in $\mathcal{G}_5$ have very the smallest caching probabilities. For example, even though f_5 is much larger than f_{11}, p_5 is smaller than p_{11} because content 11 belongs to the category of $\mathcal{G}_1$ which is the most popular among all categories.

In Fig. 2, when $L = 4$, it seems that contents in the same category have almost the identical caching probabilities, because the user is likely to request $L = 4$ contents in the same category owing to high q. Meanwhile, when $L = 2$, contents 1-5 have larger caching probabilities than other contents in the same category. In addition, as q grows, differences in caching probabilities by categories increase, i.e., the dependency of caching probabilities on the popularity model for consecutive content requests becomes larger than that for one-shot request. As α grows, i.e., the pathloss effect increases, the number of caching node candidates decreases, therefore the impact of popularity difference on caching probabilities becomes large.

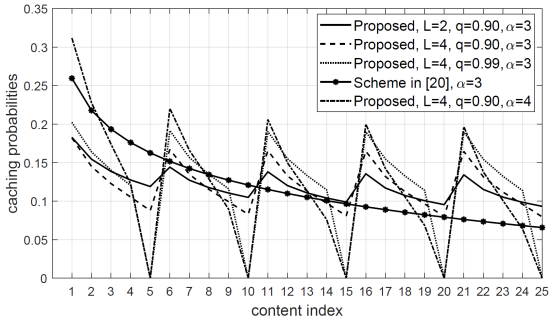

Fig. 2. Caching probabilities for each content

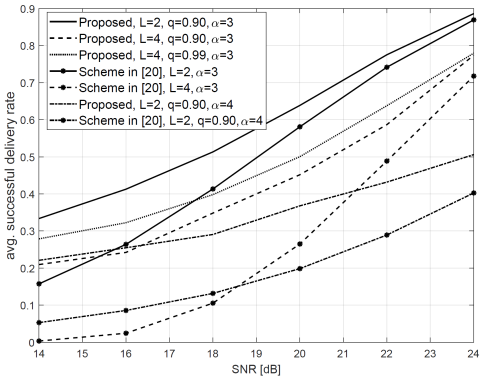

Fig. 3. The average successful delivery rate vs. SNR

In Fig. 3, the plots of average success probabilities for content delivery versus SNR are shown. Overall, the proposed caching policy outperforms the comparison scheme and the performance gain decreases as SNR grows. Additionally, a performance gain of the proposed scheme compared to the comparison technique increases as L and/or q grows. As shown in Fig. 2, caching probabilities of the proposed scheme and comparison one have more differences with larger α, therefore a performance gain of the proposed one also increases.

The impacts of the skew factor of the Zipf distribution, i.e., γ, and the PPP intensity of helpers, i.e., λ, are shown in Figs. 4 and 5, respectively. The interesting result different from [19] is that P_L^o of the comparison scheme decreases as γ, increases. Large γ makes the difference in popularity among all contents much greater. For example, f_2 becomes larger but f_{11} gets smaller as γ increases. However, as we have seen in Fig. 2, when users request multiple contents in a row, p_{11}^* is larger than p_2^* because content 11 belongs to $\mathcal{G}_1$, the most popular category. Therefore, an increase of γ does not have a good effect for the comparison scheme when the consecutive user demands for categorized contents are considered. On the other hand, P_L^o of the proposed algorithm improves as γ grows. Additionally, in Fig. 5, a performance gap between the

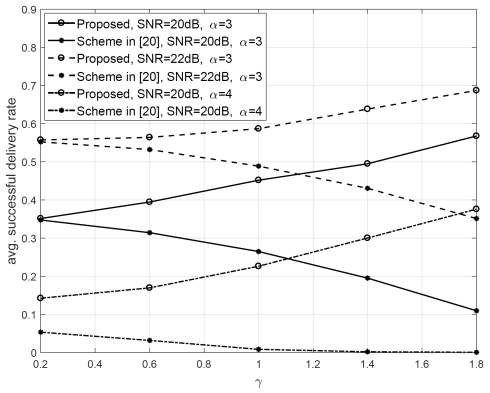

Fig. 4. The average successful delivery rate vs. γ

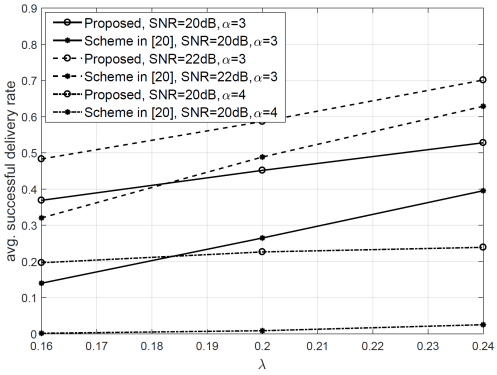

Fig. 5. The average successful delivery rate vs. λ

proposed and comparison techniques increases as the number of caching helper candidates for downloading the desired contents continuously decreases.

VI. Concluding Remarks

This paper proposes an optimal probabilistic caching policy when users request different numbers of categorized contents. The proposed scheme captures the essential characteristics of video delivery: contents in the same category have higher relevance and different users demand different content consumption. The optimal caching probabilities for multiple contents are obtained by iteratively optimizing the subproblem with respect to two contents in order to maximize the minimum of the average successful delivery rates of all users. The impacts of categorized contents and consecutive user demands on the caching policy are clearly shown by numerical results.

Acknowledgment

This work was supported by Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No.2018-0-00170, Virtual Presence in Moving Objects through 5G). Joongheon Kim is the corresponding author.

References

[1] "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 20162021 White Paper, Cisco. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/serviceprovider/visual-networking-index-vni/mobile-white-paper-c11-520862.html

[2] X. Cheng, J. Liu, and C. Dale, "Understanding the Characteristics of Internet Short Video Sharing: A YouTube-based Measurement Study," *IEEE Transactions on Multimedia*, vol. 15, no. 5, pp. 1184–1194, August 2013.

[3] J. Koo, J. Yi, J. Kim, M. A. Hoque, and S. Choi, "REQUEST: Seamless dynamic adaptive streaming over HTTP for multi-homed smartphone under resource constraints," in *Proc. ACM Multimedia*, Mountain View, CA, USA, 2017.

[4] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "FemtoCaching: Wireless Video Content Delivery through Distributed Caching Helpers," in *Proc. IEEE INFOCOM*, Orlando, FL, USA, 2012.

[5] E. Bastug, M. Bennis, and M. Debbah, "Living on the Edge: The Role of Proactive Caching in 5G Wireless Networks," *IEEE Communications Magazine*, vol. 52, no. 8, pp. 82–89, August 2014.

[6] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. C. M. Leung, "Cache in the Air: Exploiting Content Caching and Delivery Techniques for 5G Systems," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 131–139, February 2014.

[7] N. Golrezaei, A. F. Molisch, A. G. Dimakis, and G. Caire, "Femtocaching and Device-to-Device Collaboration: A New Architecture for Wireless Video Distribution," *IEEE Communications Magazine*, vol. 51, no. 4, pp. 142–149, April 2013.

[8] M. Ji, G. Caire, and A. F. Molisch, "Fundamental Limits of Caching in Wireless D2D Networks," *IEEE Transactions on Information Theory*, vol. 62, no. 2, pp. 849–869, February 2016.

[9] M. Ji, G. Caire, and A. F. Molisch, "Wireless Device-to-Device Caching Networks: Basic Principles and System Performance," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 1, pp. 176–189, January 2016.

[10] J. Kim, G. Caire, and A. F. Molisch, "Quality-Aware Streaming and Scheduling for Device-to-Device Video Delivery," *IEEE/ACM Transactions on Networking*, vol. 24, no. 4, pp. 2319–2331, August 2016.

[11] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "FemtoCaching: Wireless Content Delivery Through Distributed Caching Helpers," *IEEE Transactions on Information Theory*, vol. 59, no. 12, pp. 8402–8413, December 2013.

[12] J. Song, H. Song, and W. Choi, "Optimal Caching Placement of Caching System with Helpers," in *Proc. IEEE Int'l Conf. on Communications (ICC)*, London, UK, 2015.

[13] J. Li, Y. Chen, Z. Lin, W. Chen, B. Vucetic, and L. Hanzo, "Distributed Caching for Data Dissemination in the Downlink of Heterogeneous Networks," *IEEE Transactions on Communications*, vol. 63, no. 10, pp. 3553–3568, October 2015.

[14] C. Psomas, G. Zheng, and I. Krikidis, "Cooperative Wireless Edge Caching with Relay Selection," in *Proc. IEEE Int'l Conf. on Communications (ICC)*, Paris, France, 2017.

[15] K. Poularakis, G. Iosifidis, and L. Tassiulas, "Approximation Algorithms for Mobile Data Caching in Small Cell Networks," *IEEE Transactions on Communications*, vol. 62, no. 10, pp. 3665–3677, October 2014.

[16] W. Jiang, G. Feng, and S. Qin, "Optimal Cooperative Content Caching and Delivery Policy for Heterogeneous Cellular Networks," *IEEE Transactions on Mobile Computing*, vol. 16, no. 5, pp. 1382–1393, May 2017.

[17] B. Blaszczyszyn and A. Giovanidis, "Optimal Geographic Caching in Cellular Networks," in *Proc. IEEE Int'l Conf. on Communications (ICC)*, London, UK, 2015.

[18] Z. Chen, N. Pappas, and M. Kountouris, "Probabilistic Caching in Wireless D2D Networks: Cache Hit Optimal Versus Throughput Optimal," *IEEE Communications Letters*, vol. 21, no. 3, pp. 584–587, March 2017.

[19] S. H. Chae and W. Choi, "Caching Placement in Stochastic Wireless Caching Helper Networks: Channel Selection Diversity via Caching," *IEEE Transactions on Wireless Communications*, vol. 15, no. 10, pp. 6626–6637, October 2016.

[20] M. Choi, J. Kim, and J. Moon, "Wireless Video Caching and Dynamic Streaming Under Differentiated Quality Requirements," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 6, pp. 1245–1257, June 2018.

[21] M. Lee, A. F. Molisch, N. Sastry and A. Raman, "Individual Preference Probability Modeling for Video Content in Wireless Caching Networks," *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, Singapore, 2017, pp. 1-7.