A Diversified Recommendation Scheme for Wireless Content Caching Networks

Kexin Shi, Yaru Fu, *Member, IEEE*, and Kevin Hung, *Senior Member, IEEE*

***Abstract*—Wireless cellular networks currently face constantly growing data demands, which lead to network congestion and high latency. Cache-aware recommendation can reshape users' request behavior and improve cache efficiency in wireless caching systems, thus alleviating network congestion and shorting transmission delay. However, existing cache-aware recommendations serve the caching system by reducing the quality of recommendations. Specifically, it usually provides only limited and similar content items and lacks diversified recommendation services, which severely reduces users' satisfaction. To tackle this challenge, we propose a diversified recommendation mechanism-based solution that aims to simultaneously improve the performance of wireless caching systems and the quality of recommendation to improve users' satisfaction to a greater extent. To this end, we propose a quantitative model that captures the impact of recommendation decisions on the diversity of recommendation sets. This model enables us to formulate a joint cache hit ratio and recommendation diversity maximization problem, taking into account each user's recommendation size and cache capacity requirements. Since this problem is a non-convex integer programming problem, we decompose it into two subproblems, i.e., the cache placement problem and the diversified recommendation problem. Then we design tabu search-assisted and simulated annealing-oriented algorithms to solve these two subproblems, respectively, and perform iterative alternating optimization for the whole problem. Monte-Carlo simulation validates the effectiveness of our method in terms of cache hit ratio and recommendation diversity compared to various benchmarks.**

***Index Terms*—Cache-aware recommendation, cache hit ratio, joint optimization, recommendation diversity, time-efficient solution.**

I. Introduction

As global mobile data traffic and the proliferation of internet-of-things applications continue to surge, wireless communication systems are encountering substantial computational burdens and escalating data volumes [1]. According to Ericsson's Mobility Report, as of the first quarter of 2023, the world's monthly mobile network data traffic has reached 126 EB. It is estimated that by the end of 2028, the global mobile data traffic will reach 453 EB per month [2]. The exponential growth of mobile communication traffic poses a significant challenge for wireless communication systems. Statistics show that the growth of network traffic is mainly due to the increase in video content consumption [3]. The repeated transmission of popular content greatly increases the burden on backhaul links and increases multimedia traffic. As an illustration, Facebook encounters a staggering number of 8 billion video requests daily, with approximately 83% of these requests attributed to the top 1% most popular video content [4]. To tackle these obstacles, the utilization of wireless content caching has emerged as a promising technology, gaining significant traction in the advancement of the sixth generation of cellular networks [5]. Wireless content caching systems can be deployed to pre-cache popular files via edge facilities [6]. When a user requests content, it can be delivered directly to the user via an edge node (base station or mobile subscriber side), without traversing a complex backhaul link. This greatly alleviates network congestion, reduces the pressure on the mobile core network links, and reduces redundant data transmission [7], [8]. As an instance, in [9], authors adopted scalable video coding (SVC) to encode the content designated for caching and subsequent requests into distinct layers. Each layer set offers diverse viewing qualities, enabling the helper to cache content and thereby diminish transmission latency. Moreover, since adaptive bitrate streaming leads to further increase in video data volume, the authors in [10] optimized the adaptive video streaming transmission delay and energy consumption. Furthermore, in [11], a pioneering secure random caching scheme was introduced specifically designed for large-scale multi-antenna heterogeneous wireless networks. The proposed scheme enables base stations (BSs) to securely deliver confidential contents that are randomly cached to legitimate users while considering the presence of both passive eavesdroppers and active jammers. In addition, device-to-device (D2D) aided wireless caching networks have also received enormous attention from researchers. Popular content can be stored in the cache entity of mobile devices and shared with neighboring users through D2D communication, making a considerable portion of preferred content ubiquitous to consumers. For instance, in [12], authors studied the performance comparison between D2D caching and BS caching and found that D2D caching provides more opportunities in high-density user scenarios. In [13], authors studied how to optimize cache location in D2D networks based on user preference distribution to improve content access latency and traffic offloading gain. In [14], users were grouped into different clusters based on their preferences in D2D networks, and cache hit ratio was further optimized within each cluster. However, it is important to note that compared to the storage space of the core network, the storage space of a single BS or end user is

The research presented in this manuscript received support from the Katie Shu Sui Pui Charitable Trust through the Research Training Fellowship (Project Reference No. KSRTF / 2022 / 04). *(Corresponding author: Yaru Fu.)*

K. Shi, Y. Fu, and K. Hung are with the School of Science and Technology, Hong Kong Metropolitan University, Hong Kong, 999077, China (e-mail: s1305223@live.hkmu.edu.hk; yfu@hkmu.edu.hk; khung@hkmu.edu.hk).

restricted. The diverse distribution of user preferences and the constrained cache capacity result in insufficient fulfillment of users' needs. To address this challenge, the implementation of active mechanisms becomes crucial in enhancing cache performance.

On this ground, recommendation systems have received widespread attention for their ability to reshape user content request behavior [15]. Their objective is to deliver tailored recommendations for various types of content, such as movies, video clips, music songs, and other similar items, that align closely with the specific interests and preferences of individual users. This, in turn, increases user satisfaction and engagement, as it increases the number of clicks or content downloads. For example, Netflix's recommendation system is considered to account for 80% of clicks [16], while related video recommendations generate about 30% of views on YouTube [17]. Traditional recommendation systems tend to attract each user by recommending items that are closest to user preferences, which is not cache-friendly intuitively. Some studies have shown that joint recommendation and cache strategy design can significantly improve the performance of caching networks. As an example, in [18], the authors introduced a model that encompasses the interplay between caching decisions and user recommendations. The system is able to recommend contents that not only aligns with the user's interests but also maximizes the cache hit ratio of the wireless content caching system. In addition, authors in [19] investigated a social-aware joint recommendation and caching policy design, which can benefit the D2D-assisted wireless content caching networks. Moreover, in [20], researchers investigated the utilization of recommendations to enhance the efficiency of D2D-enabled wireless content caching networks. Furthermore, a joint recommendation and caching optimization algorithm was developed for multi-BS cooperative caching networks based on user preference prediction in [21]. Thereof, a deep crossover model is used to predict user preferences, and the result can effectively reduce the total transmission delay of the system. A multi-antenna assisted multi-cell edge network with caching-aware recommendation and user-associated transmission beamforming was investigated in [22], with the aim of minimizing the content delivery delay for mobile users. Similarly, the authors in [23] presented a scheme that combines recommendation, caching, and beamforming in multi-cell multi-antenna Fog radio access networks (Fog-RANs) with a specific emphasis on recommendation awareness.

Existing work [18]–[23] has demonstrated the effectiveness of cache-aware recommendation systems, but it is essentially using the recommendation system to serve the cache system, i.e., to improve the caching efficiency. While the performance of the wireless content caching system is improved, conversely, the recommendation system thus degrades its own quality, which potentially leads to a decrease in user satisfaction. More precisely, to improve the cache hit ratio of the wireless content caching system, the recommended items become limited and similar, reducing result diversity and limiting user choices. However, due to the heterogeneity of users and the growing diversity of their preferences, overly similar recommendation results are no longer sufficient to meet the needs of users [24]. Thus, the concept of recommendation diversity was introduced [25]–[29]. Intuitively, the goal of diversified recommendations is to identify a recommendation list that is different from the user's interests but still relevant. However, to the best of our knowledge, the issue of how to improve the quality of recommendation systems, i.e., recommendation diversity, while enhancing the caching efficiency of wireless content caching system has not been addressed, which motivates our work.

In our work, we consider improving the recommendation quality while maintaining the high cache hit ratio of the cache-aware recommendation system. The novelty of our work is that we do not sacrifice the quality of the recommendation system while improving the performance of the wireless content caching system. The result is an improvement in both the wireless content caching performance and the user experience. To achieve this goal, we introduce a metric that measures the diversity of recommendation sets per user and the performance of the wireless content caching system. Specifically, we propose a joint diversified recommendation and cache placement scheme to optimize the cache hit ratio and the diversity of recommendation sets per user for the D2D-enabled wireless content caching network.

To keep it concise, we summarize the main contributions of this study as follows:

- We consider a generic D2D communication-assisted wireless content caching network and propose two models: one is quantitatively explains how recommendation decisions affect the diversity of recommendation sets, and another one defines how recommendations and caching mathematically affect the caching efficiency of the system. Based on this, we formulate a joint cache hit ratio and recommendation diversity maximization problem, taking into account each user's recommendation size and cache capacity requirements.
- To solve this non-convex integer programming problem that jointly optimizes cache hit ratio and recommendation diversity in polynomial time, we decompose it into two subproblems and optimize them separately. First, we study the cache placement problem under a fixed recommendation policy. Second, we solve the recommendation decision problem, when the caching decision is given. Finally, we optimize these two types of variables alternately to achieve a time-efficient suboptimal solution.
- However, the foregoing mentioned subproblems are still non-convex and intractable. We design corresponding algorithms to solve them separately. Concretely speaking, we design a tabu search-assisted algorithm to solve the cache placement subproblem, and develop an simulated annealing-oriented method to solve the recommendation subproblem. Thereafter, alternating optimization is performed, and this algorithm features iterative work and provable convergence guarantees. At last, we provide a detailed analysis of the computational complexity of our devised algorithm and validate its superiority over various baselines with extensive Monte-Carlo simulations.

The rest of this work is organized as follows: Section

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2023.3343364

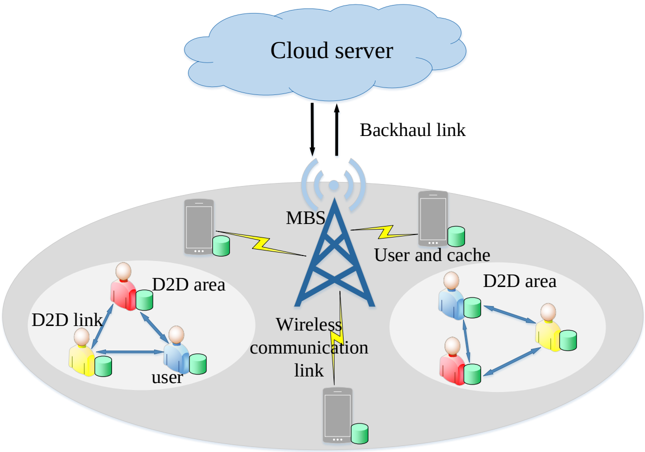

Fig. 1. An illustration of the considered D2D-enabled wireless content caching network.

II introduces the system model of our D2D-aided wireless content caching network, provides the mechanism for recommendations and the definition of recommendation diversity. Based on the overview, Section III proposes a joint recommendation and cache planning problem that maximizes both cache efficiency and recommendation diversity. Since the formulated problem is non-convex and complex, in Section IV, we decompose it into two subproblems and develop an time-efficient suboptimal algorithm to tackle it. The designed strategy is comprised by two different algorithms to solve the two subproblems separately. Numerical results are presented in Section V to demonstrate the convergence performance of our proposed algorithm using Monte-Carlo simulation, and validate its effectiveness compared to extensive benchmarking schemes. Finally, we summarize this work and point out further research directions.

II. System Model

In this section, we first introduce the model of our D2D-assisted wireless content caching network and elaborate on the caching mechanism of each user. Then we present the preference distribution of users for items, establish the recommendation mechanism, and define the content request distribution for each user within the system. Finally, we provide a detailed explanation of the metrics for recommendation diversity.

A. System Description

In our depicted scenario represented in Fig. 1, we examine a wireless content caching system with D2D assistance. The system involves a macro base station (MBS) catering to U users, who are uniformly distributed within the disk-shaped region. Denoting the index set of all users by $\mathcal{U} = \{1, 2, ..., U\}$. Our system contains I content items and define $\mathcal{I} = \{1, 2, ..., I\}$ as the index set of all items. For $i \in \mathcal{I}$, let L_i be the size of content i. Assuming that each user has a different size of cache capacity. On this basis, we define B_u as the cache capacity of user u, where $u \in \mathcal{U}$. For $u \in \mathcal{U}$ and $i \in \mathcal{I}$, we define the binary indicator $c_{u,i}$ to represent the caching decision of user u regarding content item i. Specifically, $c_{u,i} = 1$ indicates that user u has item i cached, while $c_{u,i} = 0$ indicates otherwise. It is important to note that these caching decisions are subject to the cache capacity constraint, resulting in the following condition:

$$\sum_{i \in \mathcal{I}} c_{u,i} L_i \leq B_u, \ \ u \in \mathcal{U}. \tag{1}$$

If user u requests content i, there is no requirement to establish either a D2D communication link or a wireless communication link if content i is already cached in the local storage of user u. However, if content i is not cached locally, user u will initiate a broadcast request to its D2D neighbors. In this case, user u can download content i through a D2D link if any of its D2D neighbors have it cached, avoiding the need for establishing a complex wireless channel with the MBS. In cases where none of the aforementioned conditions are met, the MBS and user u establish a wireless communication link. Upon receiving a user's request, the MBS retrieves the required content from the cache of the cloud content provider, subsequently delivering it to the corresponding user. As a result, it becomes apparent that cache-enabled D2D communication effectively alleviates the workload on the backhaul link. Moreover, we introduce D_u as the collection of all users situated within the D2D communication region of user u, where $u \in \mathcal{U}$.

B. Modeling for the Content Preferences of Users

In this section, we elaborate on the content preference of users, which is composed by two components, i.e., the feature vector of items and the feature vector of users. [1] Let $\boldsymbol{f}_i = (f_i(1), f_i(2), \ldots, f_i(M))$ be the feature vector of content i, wherein $f_i(j) \in [0, 1]$ depicts the similarity of content i in terms of feature j and M denotes the total number of features. Thereof, the features can be taken as different categories of all items. Moreover, we have:

$$\sum_{j=1}^{M} f_i(j) = 1, \ \ i \in \mathcal{I}. \tag{2}$$

Similarly, we can obtain the feature vector of user u, which is denoted by $\boldsymbol{g}_u = (g_u(1), g_u(2), ..., g_u(M))$, wherein $g_u(j) \in [0, 1]$ represents the extent to which user u likes the content under topic category j. We also normalize these values as follows:

$$\sum_{j=1}^{M} g_u(j) = 1, \ \ u \in \mathcal{U}. \tag{3}$$

For the purpose of this study, we consider the feature vectors $\boldsymbol{f}_i$ and $\boldsymbol{g}_u$ as pre-existing information since content-providing sites can estimate them based on information in regard to user's content requests and download histories [31]–[33]. Due to the disparities among users, we make the assumption that each user possesses distinct personalized preferences for each content item i. For $u \in \mathcal{U}$, $i \in \mathcal{I}$, we define $a_{u,i}^{\text{pref}} \in [0, 1]$ as user u's preference for item i, i.e., the probability of

[1] In this paper, the feature vectors for users and contents are assumed to be known, which can be predicted by machine learning-driven algorithms [30].

user u's inherent content request for item i. The collective representation of user u's inherent personal preference for item i can be established by considering the feature vectors $\boldsymbol{f}_i$ and $\boldsymbol{g}_u$ together. In this study, we adopt the cosine similarity index to measure the similarity between these two vectors, expressed as follows:

$$\widetilde{a}^{\mathrm{pref}}_{u,i} = \frac{\sum_{j=1}^{M} g_u(j) f_i(j)}{\sqrt{\sum_{j=1}^{M}(g_u(j))^2}\sqrt{\sum_{j=1}^{M}(f_i(j))^2}},\ u \in \mathcal{U},\ i \in \mathcal{I}. \tag{4}$$

For user u, we normalize all the items in $\mathcal{I}$ to obtain the preference distribution of user u in terms of content i, which is expressed as follows:

$$a^{\mathrm{pref}}_{u,i} = \frac{\widetilde{a}^{\mathrm{pref}}_{u,i}}{\sum_{i=1}^{\mathcal{I}} \widetilde{a}^{\mathrm{pref}}_{u,i}},\ i \in \mathcal{I}. \tag{5}$$

In the absence of recommendations, every user exhibits an inherent preference distribution for the content items in $\mathcal{I}$, referred to as $a^{\mathrm{pref}}_{u,i}$, where $u \in \mathcal{U}$. However, when recommendations are incorporated, each user's content requests are influenced by both their inherent preferences and the recommendation mechanism. We will provide a comprehensive description of this effect in the subsequent subsection.

C. The Impact of Recommendation on User's Content Request

In systems that incorporate recommendations, the likelihood of content requests from each user is influenced jointly by their inherent preference distribution and the recommendation strategy. Generally, recommended items exhibit an increased probability of being requested, while the request probability of non-recommended items tends to decrease. In real-world scenarios, users may choose to reject recommendations. To capture this aspect, the probability of user u accepting a recommendation is denoted by x_u, which lies within the range of $[0, 1]$. Conversely, the probability of user u rejecting a recommendation is expressed as $(1 - x_u)$, also falling within the range of $[0, 1]$. In this particular study, it is assumed that x_u is known beforehand and varies across users due to the heterogeneity of their personalities [18].

For $u \in \mathcal{U}$, $i \in \mathcal{I}$, we define $r_{u,i} \in \{0, 1\}$ as a binary indicator to show whether content i is recommended to user u. If item i is recommended to user u, $r_{u,i} = 1$, otherwise 0. In addition, we set a constraint on the number of recommendations per user. Let $\mathcal{R}_u$ be the index set of all items recommended to user u. In addition, define R_u as the maximum number of items recommended to user u, where $u \in \mathcal{U}$. With the definitions, we have:

$$\mathcal{R}_u = \{i | r_{u,i} = 1, i \in \mathcal{I}\},\ u \in \mathcal{U}. \tag{6}$$

We denote $a^{\mathrm{req}}_{u,i}$ as the content request probability for user u. To be more precise, the request probability of user u for a recommended item $i \in \mathcal{R}_u$ is formulated as follows:

$$\widehat{a}^{\mathrm{req}}_{u,i} = \frac{r_{u,i} a^{\mathrm{pref}}_{u,i}}{\sum_{j \in \mathcal{I}} r_{u,j} a^{\mathrm{pref}}_{u,j}},\ u \in \mathcal{U},\ i \in \mathcal{I}. \tag{7}$$

Meanwhile, the request probability of user u for non-recommended item $i \in \mathcal{I} \backslash \mathcal{R}_u$ is quoted as follows:

$$\widetilde{a}^{\mathrm{req}}_{u,i} = \frac{(1 - r_{u,i}) a^{\mathrm{pref}}_{u,i}}{\sum_{j \in \mathcal{I}} (1 - r_{u,j}) a^{\mathrm{pref}}_{u,j}},\ u \in \mathcal{U},\ i \in \mathcal{I}. \tag{8}$$

By amalgamating the aforementioned two analyses into a consolidated formula, we derive the request probability of user u for item i in the following manner:

$$a^{\mathrm{req}}_{u,i} = x_u \widehat{a}^{\mathrm{req}}_{u,i} + (1 - x_u) \widetilde{a}^{\mathrm{req}}_{u,i},\ u \in \mathcal{U},\ i \in \mathcal{I}. \tag{9}$$

It is evident that for $i \in \mathcal{I}$, the content request distribution of user $a^{\mathrm{req}}_{u,i}$ maintains its normalization property.

D. Recommendation Diversity

In conventional recommendation system, each user will be recommended by a set of its top-preferred items. These items are likely to be very similar. Thereby, users may get bored with the recommendation results [34]–[36]. To address this issue, we should consider both the relevance and the diversity properties of recommendation set, under which a combined optimized utility function is proposed. This combined optimization utility function consists of a modular function and a dispersion function. We use the modular function to model each item's relevant properties, and the dispersion function describes the diverse properties of an item set. More specifically, we define $R_m(\mathcal{R}_u)$ as the summation of the m-th category's relevance properties for of all items in the recommendation set $\mathcal{R}_u$ for user u. According to the previous section, the relevance properties is defined as the different categories of items, i.e.,$\boldsymbol{f}_i$, where $i \in \mathcal{I}$. On this basis, the modular function is expressed as follows:

$$R_m(\mathcal{R}_u) = \sum_{i \in \mathcal{R}_u} f_i(m), u \in \mathcal{U}. \tag{10}$$

Similarly, we let $V(\mathcal{R}_u)$ be the dispersion function that captures the diversified properties of recommendation set $\mathcal{R}_u$ [29]. For fair comparison, we follow [37] to use the average Cosine distance of all pairs of items in the recommendation set as the distance metric of the dispersion function, which is defined as follows:

$$V(\mathcal{R}_u) = \frac{2}{|\mathcal{R}_u|(|\mathcal{R}_u| - 1)} \sum_{\{i,j\}: i,j \in \mathcal{R}_u} [1 - \mathrm{sim}(\boldsymbol{f}_i, \boldsymbol{f}_j)], u \in \mathcal{U}, \tag{11}$$

where $\mathrm{sim}(\boldsymbol{f}_i, \boldsymbol{f}_j)$ denotes the Cosine similarity between the feature vectors of two items i and j in $\mathcal{R}_u$.

With foregoing analysis, we define the user preferences about relevance and diversity as $\eta = [\boldsymbol{g}_u^T, \beta_u^T]^T$, which belong to the vector space $\mathbb{R}^M$ and are prior information. Our goal is to find a set $\mathcal{R}_u \in \mathcal{I}$ that maximizes the following utility function:

$$F(\mathcal{R}_u | \eta) = \sum_{m=1}^{M} g_u(m) R_m(\mathcal{R}_u) + \beta_u V(\mathcal{R}_u), u \in \mathcal{U}, \tag{12}$$

where $R_m(\mathcal{R}_u)$ is well defined in (10). $g_u(m)$ is user u's preference for the m-th category's relevance property, where $m \in \mathbb{R}^M$. $V(\mathcal{R}_u)$ is given in (11), and β_u is the user's preference for the diversity properties for user u [29].

III. The Joint Caching and Diversified Recommendation Problem

In this section, we formulate the joint recommendation and cache placement problem for D2D-oriented wireless content caching networks, aiming to maximize system's cache hit ratio and recommendation diversity simultaneously. For simplicity, we let $\boldsymbol{c}_u = (c_{u,1}, c_{u,2}, ..., c_{u,I})$ and $\boldsymbol{c} = (\boldsymbol{c}_1, \boldsymbol{c}_2, ..., \boldsymbol{c}_u)$ be the caching strategy of user u and the cache decision vector of the network, respectively. In addition, we define $\chi(x)$ as an indicator function which is express as follows:

$$\chi(x) = \begin{cases} 1 & \text{if } x > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

Therefore, $\chi(\sum_{u \in D_u} c_{u,i})$ can be used to indicate whether the neighborhood users in the D2D region of user u have cached item i.

With aforementioned definitions, the considered joint optimization problem is formulated as follows:

$$\max_{\mathcal{R}_u,\, \boldsymbol{c}} \quad \sum_{u \in \mathcal{U}} F(\mathcal{R}_u|\eta) + \sum_{u \in \mathcal{U}} \sum_{i \in \mathcal{I}} a_{u,i}^{\text{req}} [\chi(\sum_{u \in D_u} c_{u,i})] \tag{14}$$

s.t.

$$\begin{aligned} &C_1 : \sum_{i \in \mathcal{I}} r_{u,i} = R_u, u \in \mathcal{U}, \\ &C_2 : \sum_{i \in \mathcal{I}} c_{u,i} L_i \leq B_u, u \in \mathcal{U}, \\ &C_3 : c_{u,i} \in \{0, 1\}, u \in \mathcal{U}, i \in \mathcal{I}, \\ &C_4 : r_{u,i} \in \{0, 1\}, u \in \mathcal{U}, i \in \mathcal{I}, \end{aligned}$$

$F(\mathcal{R}_u|\eta)$ and $a_{u,i}^{\text{req}}$ in the objective function are defined in (12) and (9), respectively. C_1 shows the constraint on the number of recommendations per user. C_2 reflects the cache storage capacity constraint. C_3 and C_4 describe the binary properties of the decision variables.

It is important to note that the joint optimization problem (14) is categorized as a non-convex integer programming problem. The non-convexity arises from the characteristics of the functions $\chi(\sum_{u \in D_u} c_{u,i})$ and $a_{u,i}^{\text{req}}$ with respect to the binary variables $\boldsymbol{c}$ and $\mathcal{R}_u$ respectively. Consequently, the overall objective function becomes non-convex, posing challenges in obtaining the optimal solution. Given that the cache decision $\boldsymbol{c}$ and the recommendation decision $\mathcal{R}_u$ are mutually affected with each other, in the next section, we decompose this intractable non-convex maximization problem into two subproblems to optimize the caching decision variable and the recommendation decision variable separately. Lastly, we execute alternating optimization on these two subproblems iteratively until reaching convergence.

IV. The Joint Caching and Diversified Recommendation Decision Algorithm

In this section, we study the joint cache placement and diversified recommendation problem proposed in the previous section. In order to reduce the complexity of the algorithm, we decompose the problem into two subproblems and develop the corresponding algorithms to solved them separately. Specifically, we will first discuss the cache placement problem under a given recommendation policy, which is optimized using a tabu search-assisted algorithm. Then, we employ an simulated annealing-oriented algorithm to find diversified recommendation decisions under a given cache placement decision. Finally, alternating optimization is performed iteratively to maximize the objective function while ensuring convergence.

A. Optimization of Caching Decision

In this section, we optimize the cache placement decision problem under a given recommendation policy. Thus, the joint optimization problem (14) is simplified to the following problem:

$$\max_{\boldsymbol{c}} \quad \sum_{u \in \mathcal{U}} F(\mathcal{R}_u|\eta) + \sum_{u \in \mathcal{U}} \sum_{i \in \mathcal{I}} a_{u,i}^{\text{req}} [\chi(\sum_{u \in D_u} c_{u,i})] \tag{15}$$

s.t.

$$\begin{aligned} &C_2 : \sum_{i \in \mathcal{I}} c_{u,i} L_i \leq B_u, u \in \mathcal{U}, \\ &C_3 : c_{u,i} \in \{0, 1\}, u \in \mathcal{U}, i \in \mathcal{I}. \end{aligned}$$

In the above objective function, since $\mathcal{R}_u$ is already given, $F(\mathcal{R}_u|\eta)$ and $a_{u,i}^{\text{req}}$ are known in accordance with (12) and (9), respectively. However, since the function $\chi(\sum_{u \in D_u} c_{u,i})$ containing $c_{u,i}$ is non-convex, this cache placement problem is still non-convex. Therefore, we plan to optimize it with a tabu search-assisted algorithm based on the idea about two side-exchange. Specifically, in our proposed tabu search-assisted algorithm, we search the neighbourhood with the two-side exchange algorithm on the basis of the current solution to produce a new solution. Tabu list is used to judge whether to accept this new solution or not, thus prevent falling into a local optimal solution to some extent.

Prior to delving into the intricacies of the proposed algorithm, it is necessary to establish certain definitions. We introduce $\mathcal{E}_u$ as the index set representing the cached items at user u. Notably, $\mathcal{E}_u$ can be associated with the cache decision vector of user u, denoted as $\boldsymbol{c}_u$. To simplify the notation, we define the mapping Υ: $\mathcal{E}_u \to \boldsymbol{c}_u$ as follows:

$$\Upsilon(\mathcal{E}_u) = \boldsymbol{c}_u = (c_{u,1}, c_{u,2}, ..., c_{u,I}). \tag{16}$$

Using the definitions provided earlier, we denote $\mathcal{E} = \{\mathcal{E}_u | u \in \mathcal{U}\}$ as the caching policy for the entire system. Similarly, we introduce another mapping Λ: $\mathcal{E} \to \boldsymbol{c}$ defined as follows:

$$\Lambda(\mathcal{E}) = \boldsymbol{c} = (\boldsymbol{c}_1, \boldsymbol{c}_2, ..., \boldsymbol{c}_I). \tag{17}$$

Furthermore, we define $\mathcal{V}_u = \mathcal{I} \ / \ \mathcal{E}_u$ as the complementary set of $\mathcal{E}_u$. The concept of two-side exchange originates from matching theory, which is denoted by (i, j). To explore the search space and generate a new solution, we exchange the items i and j from sets $\mathcal{E}_u$ and $\mathcal{V}_u$, respectively, which is quoted below:

$$\mathcal{E}'_u = \mathcal{E}_u \setminus \{i\} \cup \{j\}, i \in \mathcal{E}_u, j \in \mathcal{V}_u, u \in \mathcal{U}. \tag{18}$$

Tabu Search (TS) is one of AI search optimization techniques proposed by Glover based on neighborhood search [38]. The core mechanism is to prevent falling into the local optimal by creating a tabu list. In our proposed tabu search-assisted scheme, we recording the reverse operations of recently exchanged items in the tabu list. More specifically, when the exchange pair (i, j) is approved, its reverse exchange pair (j, i) is added to the tabu list to prevent returning to the original local optimal solution within a certain period of time. With foregoing discussions, the tabu list is denoted by $T^{\dagger}$, which indicates the collection of item pairs that are prohibited from being exchanged and defined as follows:

$$T^{\dagger} = \{(j, i) | i \in \mathcal{E}_u, j \in \mathcal{V}_u\}. \tag{19}$$

In our proposed tabu search-assisted algorithm, when an exchange pair is recorded in the tabu list, the pair of items will not be approved for exchange, unless it satisfies the $Aspiration\ Criterion$ [38]. Thereof, $Aspiration\ Criterion$ is the condition under which an action in the tabu list is unbanned. In this paper, we define the $Aspiration\ Criterion$ as: the new solution $\mathcal{E}'_u$ is better than the historical optimal solution. To better illustrate, we define the objective value of the caching decision $\boldsymbol{c}$ as $\Psi(\boldsymbol{c})$, and we set the historically optimal solution as $\boldsymbol{c}_{\text{best}}$. With aforementioned discussions, we specify the definition of the $Aspiration\ Criterion$ as follows:

Definition 1. *Given a pair of items (i, j) in the tabu list, where $i \in \mathcal{E}_u$ and $j \in \mathcal{V}_u$, respectively, it can be approved to be exchanged under the following condition:*

$$\Psi(\boldsymbol{c}') > \Psi(\boldsymbol{c}_{\text{best}}). \tag{20}$$

It is important to note that the length of the tabu list is fixed, which is set to $|T|$. After $|T|$ times iterations, the earliest exchange pair in the tabu list will be removed and allowing to be exchanged again. For simplicity, we summarize the details of our proposed tabu search-assisted cache placement scheme in Algorithm 1.

Algorithm 1 Cache placement algorithm

```
Considering the pre-determined recommendation strategy $\mathcal{R}_u$
  and the initial cache decision vector $\boldsymbol{c}(0)$, we proceed with the
  following settings: set the initial iteration number as $k = 0$ and
  define the maximum number of iterations as $K$; Initialize the
  tabu list as $T^{\dagger} = \emptyset$ with a specified length $|T|$; Additionally,
  denote by $\mathcal{U}^{\dagger} = \mathcal{U}$.
repeat
  for $k = 1$ to $K$ do
    Select a pair of items $(i, j)$ randomly, where $i \in \mathcal{E}_u$ and
      $j \in \mathcal{V}_u$. Exchange items pair to generate a new solution $\mathcal{E}'$
      based on (18).
    Update $\boldsymbol{c}'$ in accordance with (17).
    if $(i, j) \in T^{\dagger}$ then
      if $\Psi(\boldsymbol{c}') > \Psi(\boldsymbol{c}_{\text{best}})$ then
        Release the tabu exchange pairs $(i, j)$, i.e., $T^{\ddagger} = T^{\dagger} \setminus$
          $\{(i, j)\}$.
        Update $\boldsymbol{c} = \boldsymbol{c}'$, $\boldsymbol{c}_{\text{best}} = \boldsymbol{c}'$.
        Update $T^{\ddagger}_{*} = T^{\ddagger} \cup (j, i)$, and $(j, i)$ is released from
          $T^{\ddagger}_{*}$ after $|T|$ round iterations.
      else
        Keep current cache decision $\boldsymbol{c}$.
      end if
    else
      if $\Psi(\boldsymbol{c}') > \Psi(\boldsymbol{c}_{\text{best}})$ then
        Update $\boldsymbol{c}_{\text{best}} = \boldsymbol{c}'$.
        Update $\boldsymbol{c} = \boldsymbol{c}'$, $T^{\ddagger} = T^{\dagger} \cup (j, i)$.
      else
        Update $\boldsymbol{c} = \boldsymbol{c}'$, $T^{\ddagger} = T^{\dagger} \cup (j, i)$.
      end if
    end if
  end for
  Denote $\mathcal{U}^{\dagger} \triangleq \mathcal{U}^{\dagger} \setminus \{u^*\}$.
until $\mathcal{U}^{\dagger} = \emptyset$
return  the caching strategy $\boldsymbol{c}_{\text{best}}$.
```

B. Optimization of Diversified Recommendation Decision

In this subsection, we study the optimization of diversified recommendation decisions under a fixed caching policy $\boldsymbol{c}$. Given a cache policy $\boldsymbol{c}$, the original optimization problem (14) can be expressed as:

$$\max_{\mathcal{R}_u} \quad \sum_{u \in \mathcal{U}} F(\mathcal{R}_u | \eta) + \sum_{u \in \mathcal{U}} \sum_{i \in \mathcal{I}} a^{\text{req}}_{u,i} [\chi(\sum_{u \in D_u} c_{u,i})] \tag{21}$$

s.t.

$$C_1 : \sum_{i \in \mathcal{I}} r_{u,i} = R_u, u \in \mathcal{U},$$

$$C_4 : r_{u,i} \in \{0, 1\}, u \in \mathcal{U}, i \in \mathcal{I}.$$

Notably, given the predetermined value of $\boldsymbol{c}$, the indicator function $\chi(c)$ is established and can be interpreted as the availability of item i to user u through D2D-enabled edge caching. By referring to the content request probability definition in (9), the optimization of recommendations for user u operates independently from the strategies of other users. Consequently, the recommendation decision for all users can be divided into U parallel subproblems. For the sake of clarity, we consider user u as an illustrative example. Utilizing equations (9) and (12), it becomes evident that the objective function (21) exhibits a non-convex nature in terms of the recommendation set $\mathcal{R}_u$.

To solve this problem, we propose a simulate annealing-oriented algorithm and give the following definitions: define $\mathcal{Z}_u = \mathcal{I} \setminus \mathcal{R}_u$ as the complementary set of recommendation set $\mathcal{R}_u$ of user u. The initial solution $\mathcal{R}_u(0)$ is set by top preferred items of user u. Similar to the previous subsection, we use two-side exchange algorithm to generate new solutions $\mathcal{R}'_u$, which is quoted below:

$$\mathcal{R}'_u = \mathcal{R}_u \setminus \{i_u\} \cup \{j_u\}, i_u \in \mathcal{R}_u, j_u \in \mathcal{Z}_u, u \in \mathcal{U}. \tag{22}$$

With foregoing definitions, we use the simulated annealing algorithm to determine whether to accept the new solution generated by the two-side exchange algorithm. We define the value of objective function (21) under the recommendation decision $\mathcal{R}_u$ as $\Phi(\mathcal{R}_u)$. In our proposed simulated annealing-oriented algorithm, it has a certain probability of accepting a poor solution to prevent falling into a local optimum, which we defined it as $\mathcal{P}$. This probability is affected by the current temperature, denoted as t_k in the k-th iteration. We give the definition of the probability $\mathcal{P}$ as follows:

$$\mathcal{P} = e^{-\frac{|\Phi(\mathcal{R}'_u) - \Phi(\mathcal{R}_u)|}{t_k}}, \tag{23}$$

where $|\Phi(\mathcal{R}'_u) - \Phi(\mathcal{R}_u)|$ denotes the difference value of objective function (21) between the new solution and the current solution.

In our proposed simulated annealing-oriented algorithm, in addition to accepting a superior solution, it has a certain probability of accepting a poor solution to prevent falling into a local optimum, which we define as $Metropolis\ Criterion$ [39]. This probability is affected by the current temperature, denoted as t_k in the k-th iteration, we give the following definition:

Definition 2. *The new solution $\mathcal{R}'_u$ is accepted as the current solution when the following conditions are satisfied:*

$$\Phi(\mathcal{R}'_u) > \Phi(\mathcal{R}_u) \text{ or } \mathrm{random}(0,1) \leq \mathcal{P}. \tag{24}$$

Furthermore, in our proposed approach, we perform a certain number of two-side exchange of items at each temperature. More specifically, we let Markov chain length (referred to as J) be the number of iterations performed at each temperature [40]. After completing a full iteration of the Markov chain length, the temperature is reduced by a decay factor a. The cooling function can be expressed as follows:

$$t_{k+1} = a t_k, \tag{25}$$

where a is the annealing rate of the cooling function.

For the sake of brevity, Algorithm 2 summarizes the recommendation decision method we proposed for user u, wherein the pseudo-code representation is elaborated.

Algorithm 2 Recommendation decision algorithm for user u

```
Considering the caching placement strategy denoted as $\boldsymbol{c}$ and the initial recommendations set $\mathcal{R}_u(0)$, we proceed with the following settings: set the Markov chain length as $J$; Specify the initial temperature as $t_0$ and the minimum temperature as $t_{\min}$; Initialize the temperature interaction number as $k = 0$.
repeat
    for $n = 1$ to $J$ do
        Select a pair of items $(i_u, j_u)$, where $i_u \in \mathcal{R}_u$ and $j_u \in \mathcal{Z}_u, u \in \mathcal{U}$. Generate a new solution $\mathcal{R}'_u$ in accordance with (22).
        if $\Phi(\mathcal{R}'_u) > \Phi(\mathcal{R}_u)$ or $\mathrm{random}(0,1) \leq \mathcal{P}$ then
            Update $\mathcal{R}_u = \mathcal{R}'_u$.
        else
            Keep the current recommendation pattern.
        end if
    end for
    Lowering the temperature: $t_{k+1} = a t_k$.
    Update $k = k + 1$.
until $t_k \leq t_{\min}$
return the recommendation decision of user $u$, i.e., $\mathcal{R}_u$.
```

C. The Joint Optimization for Cache Placement and Diversified Recommendation Decision

Within this subsection, we delve into the joint optimization problem involving the cache placement algorithm and the diversified recommendation decision algorithm mentioned earlier. Additionally, we examine the convergence properties and complexity of our proposed method.

Before discussing the algorithm, some definitions are given. For $u \in \mathcal{U}, i \in \mathcal{I}$, $c_{u,1}(k)$ is defined as the caching decision of user u for item i in the k-th iteration. For brevity, we use $\boldsymbol{c}_u(k) = (c_{u,1}(k), c_{u,2}(k), ..., c_{u,I}(k))$ to denote the caching policy of user u in iteration k. Similarly, let $\boldsymbol{c}(k) = (\boldsymbol{c}_1(k), \boldsymbol{c}_2(k), ..., \boldsymbol{c}_U(k))$ be the caching policy of our method in the k-th iteration. In addition, assume that $\boldsymbol{c}(0)$ is the initial cache decision according to the top-caching scheme. Similarly, we define $\mathcal{R}_u(k)$ as the recommendation strategy in the k-th iteration, and set $\mathcal{R}_u(0)$ by the topological recommendation policy described above.

According to the definitions, the steps of the joint optimization algorithm is expressed as follows: in the k-th iteration, we start with the operation of Algorithm 1. The initial cache policy is set to $\boldsymbol{c}(k-1)$ and obtain the optimized cache policy $\boldsymbol{c}(k)$ through Algorithm 1, where fixed recommendation strategy $\mathcal{R}_u(k-1)$ is obtained from the previous iteration. Then, we using the cache policy $\boldsymbol{c}(k)$ which obtained from the previous step to determine the updated recommendation strategy $\mathcal{R}_u(k)$ by Algorithm 2, when the initial recommendation strategy is set to $\mathcal{R}_u(k-1)$. This idea ensures that the optimization strategy can produce a monotonically increasing objective value. We repeat the above steps until the value of the objective function (14) can no longer increase or the maximum number of iterations is reached. For the sake of brevity, we summarize the associated pseudo-code of our developed joint optimization method in Algorithm 3.

Algorithm 3 Joint cache decision and diversified recommendation optimization algorithm

```
Define the maximum number of iterations as $K_{\max}$. $\mathcal{R}_u(0)$ and $\boldsymbol{c}(0)$ are determined by the top-recommendation policy and the top-cache policy, respectively. Let $k = 1$.
repeat
    Running Algorithm 1: The initial caching decision is set to $\boldsymbol{c}(k-1)$, and the fixed recommendation policy is set to $\mathcal{R}_u(k-1)$. Calculate the optimized cache policy $\boldsymbol{c}(k)$.
    Running Algorithm 2: The initial recommendation decision is set to $\mathcal{R}_u(k-1)$. The fixed caching policy $\boldsymbol{c}(k)$ is obtained in the previous step. Update new recommendation policy $\mathcal{R}_u(k)$.

    Update $k = k + 1$.
until the objective function (14) cannot be further increased or the maximum iteration number $K_{\max}$ is reached.
return the joint cache placement and diversified recommendation decision strategy $(\mathcal{R}_u,\ \boldsymbol{c})$
```

Lemma 1. *The worst-case computational complexity per iteration of Algorithm 3 is max$\{\mathcal{O}(UKI^2), \mathcal{O}(NJUIR_u)\}$.*

Proof. Since Algorithm 3 works by alternating iterations of Algorithm 1 and Algorithm 2, the time complexity of Algorithm 3 depends mainly on Algorithm 1 and Algorithm 2. On the one hand, since there have U users with K iterations per user, and the complexity is related to the number of two-side exchanges. Therefore, the time complexity of Algorithm 1 is $\mathcal{O}(UKI^2)$. On the other hand, the complexity of Algorithm 2 is $\mathcal{O}(NJUIR_u)$ [21], where N is the total number of temperature changes, J is the Markov chain length, and each user will iterate a total of $N \times J$ times. Because of the two-side exchange operation within the algorithm, the complexity is related to the recommendation size R_u and the total number

TABLE I: System parameters setting

System parameters	**Values**
No. of users U	10
No. of items I	30
Users feature vector $g_u(j)$	$[0, 1]$
Items feature vector $f_i(j)$	$[0, 1]$
Network cell radius	250 m
D2D communication radius	50 m
Categories of items M	10
Item data size L_i	1
Recommendation accept ratio X_u	0.618
User preference of diversity β_u	$[0, 0.2]^{10}$
Tabu length $\lvert T \rvert$	100
The initial temperature t_0	500
The minimum temperature $t_{\min}$	1
Markov chain length J	$10R_u$

of items I [20]. □

Lemma 2. *The convergence of Algorithm 3 is warranted.*

Proof. It can be observed from Algorithm 3 that the objective value of (14) increases with the number of iterations during the iterative process and upper bounded by U and R_u. It's a monotonically increasing and upper bounded function which ensures the convergence of the joint optimization algorithm. □

Before concluding this section, it is important to acknowledge that the optimization problem involving coupled binary variables on the system objective function (14) is exceptionally challenging to analyze. A commonly adopted approach is to decompose the variables into subproblems, allowing for the optimization of each individual variable. It is important to note that our caching decision subproblem and recommendation decision subproblem remain non-convex nonlinear integer programming problems, posing significant challenges in obtaining a global optimal solution within a reasonable timeframe. To address this, we have devised tabu search-assisted and simulated annealing-oriented algorithms to tackle the caching and recommendation subproblems, respectively. Subsequently, an alternating optimization scheme is employed to iteratively improve the system objective function, ensuring a non-decreasing value from any initial state. This approach enhances the convergence performance of our joint decision-making algorithm, ultimately leading to a suboptimal solution within polynomial time. Importantly, even under worst-case conditions, the computational complexity of the developed solutions remains tolerable, making them well-suited for practical applications.

V. Numerical Results

In this section, we employ a Monte-Carlo simulation approach to showcase the effectiveness of our proposed algorithm that combines cache decision-making with diversified recommendation techniques. As shown in Table I, we summarize the system parameters for this paper. More specifically, we consider a small-sized D2D network[2] and set the cell radius as 250 m, wherein 10 users are uniformly distributed within its disk region, i.e., U = 10. Following the settings in [31], the D2D communication radius for each user is assumed to be 50 m. In addition, we consider a scenario where there are I = 30 candidate items, each associated with M = 10 distinct themes. The feature vectors for both users and content items are generated using a random walk approach. For the sake of simplicity, we assume that all content items have the same data size, normalized to 1, denoted as L_i = 1 for $i \in \mathcal{I}$. Likewise, we assume that each user has the same size storage, i.e., $B_u = B$. In addition, we also consider that each user has an identical acceptance probability[3], namely X_u = 0.618 [20] for $u \in \mathcal{U}$. In this paper, we consider not only the relevance and diversity of the recommendation set, but also each user's preference for item relevance properties and diversity properties. The parameters regarding the joint relevance and diversity performance are also summarized below. The relevance feature of each item is the content feature vector, i.e., $\boldsymbol{f}_i$, which is randomly generated. The user's preference for the relevance properties is the user's feature vector, i.e., $\boldsymbol{g}_u$. Here, we use only one-dimensional diversity features, which is defined in (11). And the user's preference for the diversity properties β_u is randomly generated from $[0, 0.2]$ for each individual user [29].

In addition, the following baselines are considered for performance comparisons:

- **Baseline 1**: Top cache and no recommendation. This scheme operates as follows: for each user u, the cache is filled with the top preferred items based on its inherent preferences $a_{u,i}^{\text{pref}}$ until the cache storage is fully utilized. Furthermore, no recommendations are provided to users under this scheme.
- **Baseline 2**: Top cache and top recommendations. In this method, user u will cache the top preferred items according to their inherent preferences. Likewise, that user will be recommended by the top preferred items based on $a_{u,i}^{\text{pref}}$.
- **Baseline 3**: Homogeneous cache and homogeneous recommendation. Under this particular strategy, every user will store the most highly ranked items determined by the collective preferences of all users. Likewise, it will recommend the top-R_u items to each user, considering the combined preferences of all users.

This section evaluates four systematic metrics: 1) the convergence performance of the joint cache and diversified recommendation algorithm we designed, 2) the cache hit ratio of different schemes, 3) the diversified recommendation value of different strategies, 4) the joint cache and diversified recommendation objective function values of different schemes.

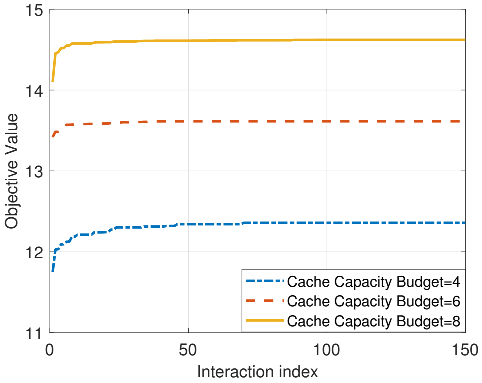

Fig. 2. Convergence behavior of our designed algorithm under different values of cache capacity budget when recommendation size is 5.

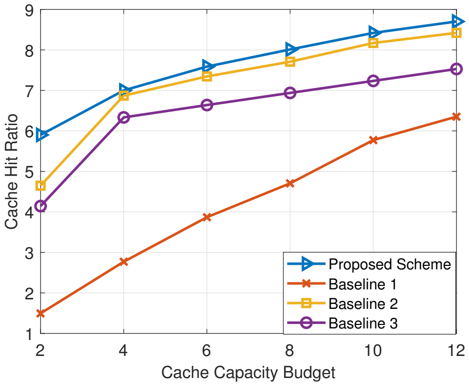

Fig. 4. Cache hit ratio versus cache capacity budget per use.

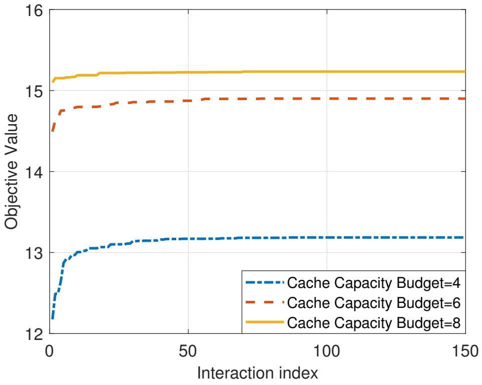

Fig. 3. Convergence behavior of our designed algorithm under different values of cache capacity budget when recommendation size is 7.

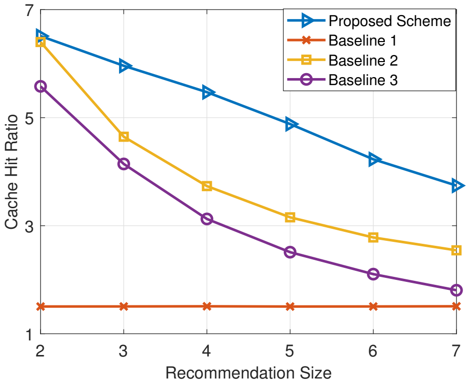

Fig. 5. Cache hit ratio versus recommendation size per user.

A. Convergence Performance

Within this subsection, we assess the convergence performance of our decision-making algorithm across various parameter configurations, as depicted in Fig. 2 and Fig. 3. We stipulate the number of recommended items per subscriber as $R_u = 5$ and $R_u = 7$ for Fig. 2 and Fig. 3, respectively. In addition, different cache capacity budget are considered for both figures. To demonstrate the convergence performance, we utilize the objective value (14) throughout the iterations. Specifically, the x-axis denotes the number of iterations, while the y-axis represents the value of the objective function.

As shown in Figs. 2 and 3, the objective value monotonically increases with the number of iterations. When comparing Fig. 2 with Fig. 3, it becomes apparent that a larger recommendation size R_u results in a higher objective value under the same cache capacity budget. Additionally, an increase in the cache capacity budget for the same recommendation size also leads to an increase in the objective value. This indicates that increasing the number of recommendations or the cache capacity budget can enhance the joint cache hit ratio and the diversity of recommendations performance to a certain extent.

B. Cache Hit Ratio

Fig. 4 and Fig. 5 illustrate the performance of our proposed algorithm in terms of cache hit ratio when compared to extensive baselines, under varying cache capacity budgets and recommendation sizes. In the following paragraphs, we distinguish the observations in these two figures. First, we consider Fig. 4, where the number of recommendation set R_u is set to 3. As expected, the cache hit ratio increases with increasing cache capacity budget for all schemes, and our designed algorithm always achieves the best performance. Noteworthy, as the cache capacity increases from 2 to 4, the gap between our scheme and baseline two is shrinks. This is due to the fact that our recommended size is set to 3, when the cache capacity is changed from 2 to 4, the cache situation changes from insufficient to sufficient. In such a case, the caching strategy based on the top-ranked will not differ much from our scheme. In addition, we can see that the gap between our solution and Baseline 2 gets larger as the cache capacity budget increases from 4. That indicates that our scheme outperforms the other baselines regardless of whether the cache capacity is smaller or lager than the recom-

[2] It is important to note that our proposed scheme is applicable to various parameter settings. Specific details are not reiterated here, as they exhibit similar performance trends.

[3] It is worth noting that the specific value chosen for the recommendation probability does not affect the performance of the algorithm. In this paper, we follow the setting described in [20] for the sake of consistency.

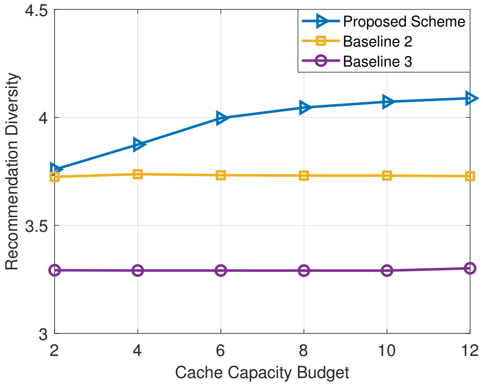

Fig. 6. Recommended diversification versus cache capacity budget per user.

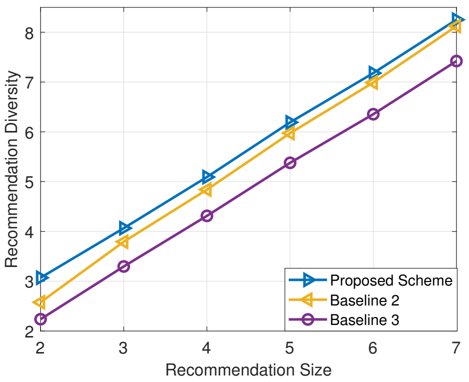

Fig. 7. Recommended diversification versus recommendation size per user.

mended number. Furthermore, Baseline 2 is the second-best performing caching strategy, and its superiority over Baseline 3 confirms the importance of considering user heterogeneity. Finally, the strategy with recommendations outperform the Baseline1 which has no recommendations, demonstrating the effectiveness of the recommendation mechanism.

We use Fig. 5 to investigate the impact of varying recommendation sizes on the cache hit ratio of all four schemes. Assuming a cache capacity budget of 2 for each user, as limited cache capacity is a prominent issue in realistic wireless edge caching systems. It can be observed that our proposed scheme outperforms all three baselines in all simulation scenarios. As shown in the figure, the cache hit ratio of all cache-aware recommendation methods decreases with an increase in the number of recommendations, which is consistent with previous research findings [20]. There are two reasons for this phenomenon. Firstly, a large number of recommendations lead to a flatter preference distribution, which fails to effectively improve cache efficiency. Secondly, an increase in the number of recommendations raises the probability of content requests, but as cache storage is limited. Too many recommendations cannot be processed, resulting in a drop in cache hit ratio. However, since our algorithm is iterative, the decrease in cache hit ratio is slower than that of other baselines. In addition, due to the consideration of user heterogeneity, the performance of Baseline 2 is superior to that of Baseline 3. Notably, the cache hit ratio of Baseline 1 remains unchanged since no recommendations are made.

C. Recommendation Diversification

Fig. 6 and Fig. 7 depict the diversity of recommendation items for the four schemes under varying cache capacity budgets and recommendation sizes. We will discuss these two figures separately. Firstly, we set the recommendation size R_u to 3 and investigate the performance changes induced by different cache capacity budgets. As shown in Fig. 6, the diversity of recommendations in our proposed algorithm increases slowly as the cache capacity budget increases. However, the diversity values of Baseline 2 and Baseline 3 remain constant. This is because the diversity of the recommendation set depends only on the choice of recommended items. Specifically, Baseline 2 and Baseline 3 always recommend the top-R_u items, and an increase in cache capacity does not affect their recommendations. Moreover, Baseline 2 consistently outperforms Baseline 3 because it considers user heterogeneity, which leads to an increase in diversity.

In the simulation shown in Fig. 7, we set the cache capacity to 8. The diversity of recommendations increases significantly and linearly with an increase in the recommendation size and our proposed algorithm consistently outperforms the other schemes. As defined in (12), we focus more on relevance than diversity because the prerequisite for increasing diversity is not to deviate from the relevance. Therefore, we set the weight of relevance much greater than that of diversity in the simulation. When the diversity of the recommendation set increases, the relevance part necessarily decreases. Thus, when the recommendation diversity utility function which defined in (12) slightly increases, it indicates a significant increase in diversity while maintaining a high level of relevance to the user's interests. Moreover, Baseline 2 still outperforms Baseline 3 in terms of diversity.

Specially, we can clearly see that the value of the recommendation diversified in (12) is directly influenced by the recommendation decision variables. Hence, for the Baseline 1 scheme, where no recommendations are made, the recommendation diversity value is 0. Consequently, the absence of Baseline 1's curve in Fig. 7 and Fig. 8 can be attributed to this fact. It's worth noting that Baseline 2 and Baseline 3 do not consider the diversity of the recommendation set or the user's preference for diversity, which may result in overly similar recommendations and unsuitable for the user.

D. System Objective Value Performance

Fig. 8 and Fig. 9 illustrate the performance comparison of our proposed joint algorithm and the other three baselines in terms of the system objective value when the cache capacity budget and recommendation size change. As shown in Fig. 8, the objective value of joint cache hit ratio and diversity of recommendations increases with an increase in the cache capacity budget, where the recommendation size is set to 3.

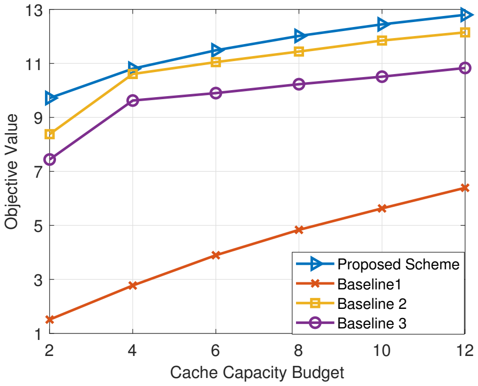

Fig. 8. The joint cache hit ratio and recommendation diversified value versus cache capacity budget per user.

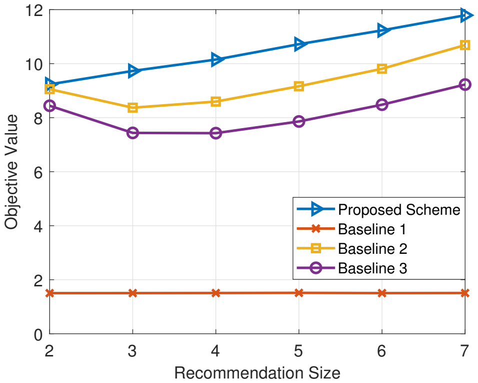

Fig. 9. The joint cache hit ratio and recommendation diversified value versus recommendation size per user.

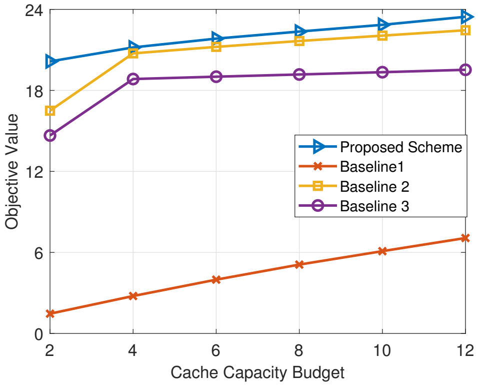

Fig. 10. The joint cache hit ratio and recommendation diversified value versus cache capacity budget per user, wherein U = 20 and I = 100.

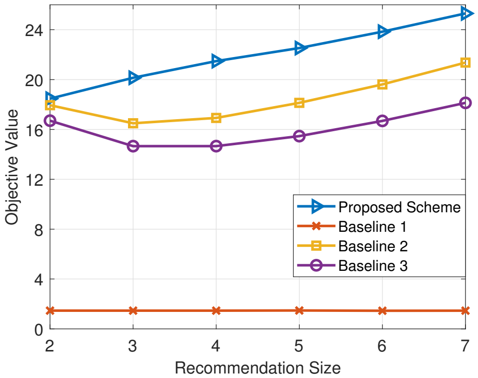

Fig. 11. The joint cache hit ratio and recommendation diversified value versus recommendation size per user, wherein U = 20 and I = 100.

Our proposed algorithm consistently outperforms the other baselines, and its growth rate is not affected by the cache capacity. In comparison, Baseline 2 and Baseline 3 experience a steep increase in their objective function value when the cache capacity budget increases from 2 to 4, followed by a gradual increase. This is because the cache capacity transitions from insufficient to sufficient during this period. Additionally, Baseline 1 does not make any recommendations, resulting in the lowest system objective function value.

In Fig. 9, the cache capacity budget is set to 2, and as the recommendation size increases, the system objective values of our proposed algorithm and Baseline 2 and Baseline 3 also increase. It's worth noting that when the recommendation size changes from 2 to 3, the system objective values of Baseline 2 and Baseline 3 experience a brief decrease, followed by an increase with the recommendation size. This is because the recommendation size exceeds the cache capacity during this stage, leading to a significant decrease in the cache hit ratio and overall system objective value. However, our proposed scheme is not affected by this special situation as our algorithm is iterative. The system objective value of Baseline 2 is better than Baseline 3, demonstrating the necessity of considering users preference heterogeneity. Additionally, the system objective value of Baseline 1 remains unchanged since it does not make any recommendations.

E. Large-scale Network Performance

In order to simulate networks with larger scales, we increase the number of users to 20 and contents to 100. The simulation of the system objective values is performed in this case, as shown in Fig. 10 and Fig. 11. We can see that the trend of our system performance in the large-scale network remains consistent with the previous section. In Fig. 10, our algorithm consistently outperforms the other baselines as the cache capacity variations gradually increase, and the joint cache hit ratio and recommendation diversified value increases as the cache capacity rises. In contrast to Fig. 8, this objective value increases as the network size increases.

In addition, Fig. 11 represents the effect of the recommendation size on the value of the system performance, consistent with the trend observed in the previous section. Baseline 2 and Baseline 3 similarly drop briefly when the recommendation size changes from 2 to 3, and then rise again as the recommendation size increases. This is because at this point the recommendation size exceeds the cache capacity, and the system cannot handle too many recommendations to fit within the cache capacity, resulting in a decrease in the cache hit ratio and the overall system objective value. Furthermore, the objective value of Baseline 1 remains constant regardless of the change in the number of users or items. This is because Baseline 1 only caches the top preferred items and does not make recommendations. Therefore, neither the cache hit ratio nor the recommendation diversity value changes.

VI. Conclusion

In this study, we focused on enhancing both the cache hit ratio and the diversity of recommendations in the cache-aware recommendation system. We aimed to optimizing the cache hit ratio of the wireless content caching system without sacrificing recommendation quality. To this end, we first quantitatively described how caching and recommendations affect cache hit ratio and recommendation diversity. Then, we considered the users recommendation size and cache capacity budget and formulated the joint maximization problem, which is a non-convex integer programming problem and difficult to find the global optimum in polynomial time. Thus, we decoupled the problem into two subproblems and designed corresponding algorithms for each subproblem. By alternating optimization these two algorithms, we obtained a joint decision-making algorithm with polynomial time complexity and guaranteed convergence. Finally, we conducted Monte-Carlo simulation, and the results showed that our algorithm achieved effective convergence and outperformed the other baselines in both cache hit ratio and recommendation diversity. In future work, we will consider the joint resource management and decision-making problems for advanced transmission techniques-oriented content caching systems with diversified recommendations. Moreover, the advanced transmission technologies encompass various approaches, including but not limited to intelligent reflecting surface (IRS) and non-orthogonal multiple access (NOMA) [41].

References

[1] W. Sun, H. Zhang, R. Wang, and Y. Zhang, "Reducing offloading latency for digital twin edge networks in 6G," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 10, pp. 12 240–12 251, Oct. 2020.

[2] Ericsson, "The Ericsson mobility report," Jun. 2023. [Online]. Available: https://www.ericsson.com/en/reports-and-papers/mobility-report.

[3] D. Wu, H. Xu, Z. Li, and R. Wang, "Video placement and delivery in edge caching networks: Analytical model and optimization scheme," *Peer-to-Peer Networking and Applications*, vol. 14, no. 6, pp. 3998–4013, Nov. 2021.

[4] K. Wang, J. Li, Y. Yang, W. Chen, and L. Hanzo, "Content-centric heterogeneous Fog networks relying on energy efficiency optimization," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 11, pp. 13 579–13 592, Nov. 2020.

[5] M. Tao, E. Chen, H. Zhou, and W. Yu, "Content-centric sparse multicast beamforming for cache-enabled cloud RAN," *IEEE Transactions on Wireless Communications*, vol. 15, no. 9, pp. 6118–6131, Sep. 2016.

[6] Q. Li, Y. Zhang, A. Pandharipande, Y. Xiao, and X. Ge, "Edge caching in wireless infostation networks: Deployment and cache content placement," *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 1–6, Apr. 2019.

[7] Z. Xiao, J. Shu, H. Jiang, J. C. S. Lui, G. Min, J. Liu, and S. Dustdar, "Multi-objective parallel task offloading and content caching in D2D-aided MEC networks," *IEEE Transactions on Mobile Computing*, pp. 1–16, Aug. 2022.

[8] T. Zhang, C. Chen, and D. Yang, "Joint user association and caching placement for cache-enabling UAV networks," *China Communications*, vol. 20, no. 6, pp. 291–309, Jun. 2023.

[9] X. Zhang, L. Zhang, Y. Ren, J. Jiang, and J. Wang, "Optimal designs of SVC-based content placement and delivery in wireless caching networks," *Sensors*, vol. 23, no. 10:4823, pp. 1–13, May. 2023.

[10] W. Liu, H. Zhang, H. Ding, and D. Yuan, "Delay and energy minimization for adaptive video streaming: A joint edge caching, computing and power allocation approach," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 9, pp. 9602–9612, Sep. 2022.

[11] W. Wen, C. Liu, Y. Fu, T. Q. S. Quek, F.-C. Zheng, and S. Jin, "Enhancing physical layer security of random caching in large-scale multi-antenna heterogeneous wireless networks," *IEEE Transactions on Information Forensics and Security*, vol. 15, no. 1, pp. 2840–2855, Dec. 2020.

[12] Z. Chen and M. Kountouris, "D2D caching vs. small cell caching: Where to cache content in a wireless network?" *2016 IEEE 17th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pp. 1–6, Aug. 2016.

[13] T. Zhang, H. Fan, J. Loo, and D. Liu, "User preference aware caching deployment for Device-to-Device caching networks," *IEEE Systems Journal*, vol. 13, no. 1, pp. 226–237, Mar. 2019.

[14] K. S. Khan, Y. Yin, and A. Jamalipour, "On the application of agglomerative hierarchical clustering for cache-assisted D2D networks," *2019 16th IEEE Annual Consumer Communications: Networking Conference (CCNC)*, pp. 1–6, Feb. 2019.

[15] Y. Zhou, J. Wu, T. H. Chan, S.-W. Ho, D.-M. Chiu, and D. Wu, "Interpreting video recommendation mechanisms by mining view count traces," *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 2153–2165, Aug. 2018.

[16] C. A. Gomez-Uribe and N. Hunt, "The Netflix recommender system: Algorithms, business value, and innovation," *ACM Transactions on Management Information Systems*, vol. 6, no. 4, pp. 1–19, Jun. 2016.

[17] R. Zhou, S. Khemmarat, and L. Gao, "The impact of YouTube recommendation system on video views," *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement*, pp. 404–410, Nov. 2010.

[18] L. E. Chatzieleftheriou, M. Karaliopoulos, and I. Koutsopoulos, "Caching-aware recommendations: Nudging user preferences towards better caching performance," *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*, pp. 1–9, May. 2017.

[19] M.-C. Lee and Y.-W. P. Hong, "Socially-aware joint recommendation and caching policy design in wireless D2D networks," *ICC 2021 - IEEE International Conference on Communications*, pp. 1–6, Aug. 2021.

[20] Y. Fu, L. Salaun, X. Yang, W. Wen, and T. Q. S. Quek, "Caching efficiency maximization for Device-to-Device communication networks: A recommend to cache approach," *IEEE Transactions on Wireless Communications*, vol. 20, no. 10, pp. 6580–6594, Oct. 2021.

[21] X. Chen, Q. Zhu, and Y. Hua, "Joint optimization of recommendation and caching based on user preference prediction," *IET Communications*, pp. 1–19, May. 2023.

[22] X. Yang, Z. Fei, B. Li, J. Zheng, and J. Guo, "Joint user association and edge caching in multi-antenna small-cell networks," *IEEE Transactions on Communications*, vol. 70, no. 6, pp. 3774–3787, Jun. 2022.

[23] X. Yang, Y. Fu, W. Wen, T. Q. S. Quek, and F. Song, "Mixed-timescale caching and beamforming in content recommendation aware Fog-RAN: A latency perspective," *IEEE Transactions on Wireless Communications*, vol. 69, no. 4, pp. 2427–2440, Apr. 2021.

[24] S. M. McNee, J. Riedl, and J. A. Konstan, "Accurate is not always good: How accuracy metrics have hurt recommender systems," *Proc.sigchi Conf.on Human Factors in Computing Systems*, pp. 1097–1101, Apr. 2006.

[25] C. Yu, L. V. S. Lakshmanan, and S. Amer-Yahia, "Recommendation diversification using explanations," *2009 IEEE 25th International Conference on Data Engineering*, pp. 1299–1302, Mar. 2009.

[26] Z. Zhang, X. Zheng, and D. D. Zeng, "A framework for diversifying recommendation lists by user interest expansion," *Knowledge-Based Systems*, vol. 105, pp. 83–95, Aug. 2016.

[27] C. Fang, H. Zhang, J. Wang, and N. Wang, "Diversified recommendation method combining topic model and random walk," *Multimedia Tools and Applications*, vol. 77, no. 4, pp. 4355–4378, Feb. 2018.

[28] L. Wang, X. Zhang, R. Wang, C. Yan, H. Kou, and L. Qi, "Diversified service recommendation with high accuracy and efficiency," *Knowledge-Based Systems*, vol. 204, pp. 1–11, Sep. 2020.

[29] Q. Ding, Y. Liu, C. Miao, F. Cheng, and H. Tang, "A hybrid bandit framework for diversified recommendation," *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, pp. 4036–4044, Dec. 2020.

[30] Y. Fu, Z. Yang, T. Q. S. Quek, and H. H. Yang, “Towards cost minimization for wireless caching networks with recommendation and uncharted users’ feature information,” *IEEE Transactions on Wireless Communications*, vol. 20, no. 10, pp. 6758–6771, 2021.

[31] B. Chen and C. Yang, “Caching policy for cache-enabled D2D communications by learning user preference,” *IEEE Transactions on Communications*, vol. 66, no. 12, pp. 6586–6601, Dec. 2018.

[32] Y. Dong, Q. Ke, Y. Cai, B. Wu, and B. Wang, “Teledata: data mining, social network analysis and statistics analysis system based on cloud computing in telecommunication industry,” *in Proceedings of the third ACM international workshop on cloud data management*, pp. 41–48, 2011.

[33] E. Bastug, M. Bennis, E. Zeydan, M. A. Kader, I. A. Karatepe, A. S. Er, and M. Debbah, “Big data meets telcos: A proactive caching perspective,” *Journal of Communications and Networks*, vol. 17, no. 6, pp. 549–557, Dec. 2015.

[34] Y. Zhang, D. Zhang, and G. Chen, “A collaborative filtering algorithm based on user preference clustering and item similarity,” *Expert Systems with Applications*, vol. 39, no. 12, pp. 10 719–10 727, 2012.

[35] N. Liu, X. Liu, and Y. Huang, “Hybrid recommendation algorithm based on user clustering and item similarity,” *Journal of Intelligent & Fuzzy Systems*, vol. 37, no. 4, pp. 4965–4975, 2019.

[36] J. Zhang, Y. Li, and X. Zhang, “A personalized recommendation algorithm based on user clustering and item similarity,” *Journal of Intelligent & Fuzzy Systems*, vol. 35, no. 1, pp. 1097–1107, 2018.

[37] N. Hurley and M. Zhang, “Novelty and diversity in top-n recommendation – analysis and evaluation,” *ACM Trans. Internet Technol*, vol. 10, no. 4, Mar 2011.

[38] F. Glover, “Tabu search for nonlinear and parametric optimization (with links to genetic algorithms),” *Discrete Applied Mathematics*, vol. 49, no. 1, pp. 231–255, 1994.

[39] N. Metropolis and S. Ulam, “The Monte Carlo method,” *Journal of the American Statistical Association*, vol. 44, no. 247, pp. 335–341, 1949.

[40] D. van Ravenzwaaij, P. Cassey, and S. D. Brown, “A simple introduction to Markov chain Monte–Carlo sampling,” *Psychonomic bulletin & review*, vol. 25, no. 1, pp. 143–154, Feb. 2018.

[41] H. Wang, C. Liu, Z. Shi, Y. Fu, and R. Song, “On power minimization for IRS-aided downlink NOMA systems,” *IEEE Wireless Communications Letters*, vol. 9, no. 11, pp. 1808–1811, Nov. 2020.

Kexin Shi received the B.E. degree from Xiamen University of Technology, Xiamen, China, in 2022, the master’s degree from the Hong Kong Metropolitan University (HKMU) in 2023. She is currently pursuing the Ph.D degree with the School of Science and Technology, Hong Kong Metropolitan University. Her research interests include wireless content caching networks, recommendation systems, resource management.

Yaru Fu (S’14-M’18) received her Ph.D in Electronic Engineering from City University of Hong Kong (CityU) in 2018. She is currently an Assistant Professor with School of Science and Technology, Hong Kong Metropolitan University (HKMU). She is presently serving as an Associate Editor for the IEEE INTERNET OF THINGS JOURNAL, the IEEE WIRELESS COMMUNICATIONS LETTERS, the IEEE NETWORKING LETTERS, and the SPRINGER NATURE COMPUTER SCIENCE. She also serves as a Review Editor for the FRONTIERS IN COMMUNICATIONS & NETWORKS and a Guest Editor for the SPACE: SCIENCE & TECHNOLOGY.

Dr. Fu was honored with the 2021 Katie Shu Sui Pui Charitable Trust - Outstanding Research Publication Award (Gold Prize), 2022 Best Editor Award for IEEE Wireless Communications Letters, 2022 Katie Shu Sui Pui Charitable Trust - Excellent Research Publication Award, and 2022 Exemplary Reviewer for the IEEE Transactions on Communications (fewer than 5%). She was listed on the World’s Top 2% Scientists 2023 ranking by Stanford University in the United States. Her research interests include intelligent wireless communications and networking, distributed storage system, and digital twin.

Kevin Hung is serving as Associate Professor and the Head of Department of Electronic Engineering and Computer Science at the School of Science and Technology, Hong Kong Metropolitan University (HKMU). He is the principal investigator of several projects funded by both the government and the university. Prior to joining HKMU, Dr. Hung was Assistant Project Manager at the Joint Research Centre for Biomedical Engineering at The Chinese University of Hong Kong (CUHK) and Engineer at a medical device company. Dr. Hung received his B.Sc. from Queen’s University in Canada, and both M.Phil. and Ph.D. from CUHK. His research interests include mobile health, wearable sensors, biosignal processing, biosystem simulation, biomedical informatics, and engineering education. In addition to his academic roles, Dr. Hung is a founding officer of the IEEE Engineering in Medicine and Biology Society (EMBS) Hong Kong – Macau Joint Chapter, and served as its Chair in 2010. He is also the founding Counsellor of the IEEE Student Branch of HKMU, the Immediate Past Chair of the Electronics and Communications Section at IET Hong Kong, a committee member of the IET Hong Kong Branch, and the Honorary Secretary of the Chinese Institute of Electronics (Hong Kong).