Q-Learning based Edge Caching Optimization for D2D Enabled Hierarchical Wireless Networks

Chenyang Wang[†‡], Shanjia Wang[†‡], Ding Li[†‡], Xiaofei Wang*[†‡] Xiuhua Li[§¶] and Victor C. M. Leung[¶]
[†]School of Computer Science and Technology, Tianjin University, Tianjin, China
[‡]Tianjin Key Laboratory of Advanced Networking, Tianjin, China
[§]School of Big Data & Software Engineering, Chongqing University, Chongqing, China
[¶]The University of British Columbia, Vancouver, Canada
*Corresponding Author: Xiaofei Wang
{chenyangwang, shanjiawang, liding_cs, xiaofeiwang}@tju.edu.cn, {lixiuhua, vleung}@ee.ubc. ca

***Abstract*—Caching at the edge of mobile networks can significantly offload network traffic while satisfying content requests from mobile users locally. The contents can be requested from the proximity users via Device-to-device (D2D) communications while proactive caching the popular content to local users. However, the assumptions that content popularity is equal to user preference in several existing studies, which are invalid and not rigorous due to the fact that content popularity is calculated by the statistic of user requests within a certain period while user preference reflects the probability of a content requested by the individual user. Motivated by this, in this paper, we study the edge caching optimization of hierarchical wireless networks. Our aiming is to maximize the size of content offload by D2D communications. In particular, the edge caching policy with D2D sharing model based on the analysis of user mobility and social relationship is derived. We first prove the problem is NP-hard and then formulate it as a Markov Decision Process (MDP) problem, finally a Q-learning based distributed content replacement strategy is proposed. The large-scale real trace based experiment results show the effectiveness of our proposed framework.**

I. Introduction

The rapid development of smart devices and wireless services have led to an explosive growth of data traffic in mobile networks. The tremendous requests for various contents (e.g., video audio stream, social networking, photos and online gaming) have posed an urgent demand for network operators to satisfy the requirements of quality of service/experience (QoS/QoE) of users towards 5G mobile networks [1]. As shown in [2], the same popular content may be downloaded many times at the similar locations. Thus, it is necessary to motivate the network operators to deploy sophisticated and advanced techniques in terms of increasing the content accessibility.

A promising technique is to cache the popular contents in proximity to the network edges (e.g., BSs and mobile devices). In this way, contents can be shared from the nearby devices to users which enables data services locally with low access latency [3] and reduces the heavy duplicated content load in backhaul networks. Besides, content sharing among mobile users via device-to-device (D2D) communications can significantly improve the spectral efficiency and reduce energy consumption [4]. Specifically, to investigate the potential gains of the hierarchical mobile edge caching architecture, content placement and delivery are two important problems which should be addressed. However, in order to design an efficient caching strategies, we need to achieve the statistical information of the user requests and sharing activities by system learning from the extreme volume of mobile traffic.

Many efforts [5] [6] have been devoted to developing the protocols and algorithms of mobile edge computing to optimize the resources efficiently. In most current studies, some key factors (such as content popularity, user preference, user behaviors, mobility patterns and spectrum usage) are assumed to be perfectly known. However, the data information of mobile users is highly spatial and temporal distributed in practice [7]. Recently, learning and big data based approaches are proposed to focus on analyzing large-scale mobile data [8] [9]. For instance, [10] explored the content caching at small base stations (SBSs) by learning the users' request interval to calculate the minimum offloading loss. Srinivasan et al. [11] used Q-learning to determine the load-based spectral to optimize the spectrum sharing. And in [12] the social behaviors of users were explored by using the decentralized deep reinforcement learning.

Wherein, several have developed cache allocation and content replacement strategies by learning the content popularity and user preference. Content popularity has significant homophily and locality [13], which is always assumed to be homogeneous in a district or among different groups. Moreover, User preference reflects the probability of a content requested by a particular user. Due to the fact that the content popularity is derived by the statistic of user requests within a certain period, which is not equal to the individual user preference, such assumptions are invalid and not rigorous.

Motivated by this, we study the optimal edge caching policy in D2D enabled hierarchical wireless networks. In particular, based on the analysis of user preference and D2D sharing activities, we derive the edge caching policy with D2D sharing model, aiming to maximize the traffic of offloaded content via D2D communications and to release the traffic pressure of Core Network. The edge caching problem is modeled by a Markov Decision Process (MDP) and the content replacement is proposed based on Q-learning to replace the contents from all the users locally with the minimum system cost.

2155-6814/18/$31.00 ©2018 IEEE
DOI 10.1109/MASS.2018.00019

There are some advantages of the proposed Q-learning based edge caching replacement strategy: 1) It is not necessary to obtain the prior knowledge (e.g. the history of user behaviors or channel state) of network, our proposed approach calculates the system cost according to the current user requests for BSs and find the best cache replacement policy; 2) The contents will be replaced during the off-peak time within each local BS, the Q-learning based approach is distributed and easy to implement; 3) Local BS exchanges the caching knowledge only with the users within its communication range, rather than the other BSs due the distributed of Q-learning, thus the system overhead cost is reduced. The contributions of this paper are summarized as follow:

- We establish the D2D sharing model by considering the social behavior and mobility of users, a hierarchical edge caching policy is proposed and the optimization problem is proved to be NP-hard.
- We model the content replacement problem as a Markov Decision Process (MDP) and deploy a distributed Q-learning based content replacement strategy.
- Integrated with the theoretical model, real trace evaluation and simulation experiment platform, the proposed Q-learning based caching scheme always outperforms the existing LRU and LFU schemes in terms of hit rate, the size of content offload and content access delay.

The rest of this paper is organized as follows. The related work is demonstrated in Sec. II. We introduce the system model in Sec. III. The details of caching policy optimization is presented in Sec. IV. We conduct the large-scale real trace based experiment in Sec. VI. Finally, we conclude the paper and envison the future work in Sec. VII.

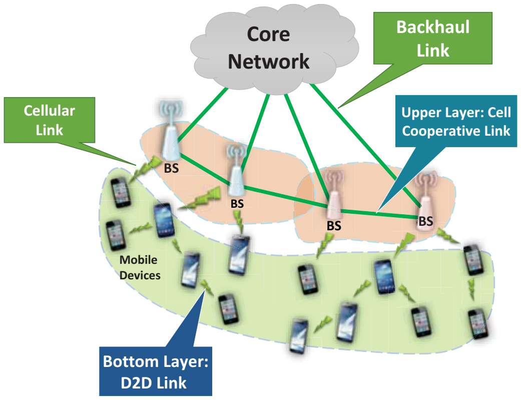


Fig. 1. Illustration of edge caching architecture in D2D networks.

II. Related Work

Several efforts are devoted on content caching at BSs in mobile networks. For instance, the potentials of caching in wireless mobile networks were explored in the surveys [14] [15]. FemtoCaching in [16] [17], and AMVS-NDN in [18] were proposed to cache the popular contents at BSs to offload the network traffic. The studies in [19] proposed the strategies of collaborative caching at BSs to improve users' QoS especially on access delay. The analysis of energy efficiency and the design of the corresponding energy-efficient caching schemes in wireless networks were discussed in [21] [22]. The authors of [23] proposed wireless content delivery schemes from BSs to users. However, these works only focused on the case of single-tier caching.

In practice, multi-tier caching has been widely used to explore the potentials of the system infrastructures especially in web caching systems [24] [25] and IPTV systems [26]. However, there are only few works utilizing the idea of multi-tier caching in mobile networks with hierarchical structures, e.g., HetNets. For instance, the survey in [26] discussed the caching paradigm in two-tier small cell networks based on user social structures. [27] focused on the theoretical performance analysis on content caching in a three-tier HeNets where content sizes were assumed to be identical. However, [27], [26] did not involve the design of caching strategies with practical considerations on some constraints (e.g., limited-capacity fronthaul/backhaul, diversity of content sizes) and specific characteristics of the network topology.

III. System Model

As shown in Fig.1, we consider a hierarchical network architecture with Core network, associated with $\mathcal{N}$ base stations (BSs) via backhaul links. N mobile users are uniformly distributed $\mathcal{U} = \{u_1, u_2, ..., u_N\}$ with a local buffer size $\mathcal{L}^u = \{l_1^u, l_2^u, ..., l_N^u\}$, users can establish direct communications with each other via D2D links,they can also be served by the BSs via cellular links. M files are stored in the content library $\mathcal{F} = \{f_1, f_2, ..., f_M\}$, and their content sizes denoted as $\mathcal{L}_f = \{l_1^f, l_2^f, ..., l_M^f\}$. The cache state is described by an $N \times M$ matrix $\Phi := (s_{uf})^{N \times M}$. Here, s_{uf} for each $u \in \mathcal{U}$ and $f \in \mathcal{F}$ is binary, where $s_{uf} = 1$ denotes the user u caches the content f while $s_{uf} = 0$ means no caching.

A. Content Popularity and User Preference

Content popularity generally describes the probability distribution of content requests in the library $\mathcal{F}$ from all the users in the network. Denote an $N \times M$ popularity matrix $\mathcal{P}$, where $q_{uf} = \mathcal{P}(q_{nm})$ is the probability of user u_n requests for content f_m in the $(n, m)^{th}$ component. In related works, the content popularity is always described by Zipf distribution as [28].

$$q_{uf} = \frac{R_{uf}^{-\beta}}{\sum_{i \in \mathcal{F}} R_i^{-\beta}}, \tag{1}$$

where the R_{uf} is popularity index that user u gives to content f in a descending order and $\beta \geq 0$ is the Zipf exponent.

Although the matrix $\mathcal{P}$ may change over time in practice, according to the measurement of users' sharing activities from $Xender$ real trace (a large- scale trace of D2D sharing, we will introduce the trace in Section VI in detail), shown as

Fig.3. Assume that the matrix stays relatively constant during a certain period and our cache policy tends to be refreshed along with the updating of popularity matrix $\mathcal{P}$. And the period of user sharing activities can be divided into Peak hours and Peak-off hours. The cache replacement action occurs during the Peak-off hours at each period.

$User\ preference$, denoted as $\mathcal{P}_{uf}$, is the probability distribution of a user's request for each content. According to the content popularity matrix $\mathcal{P}$, each row of the matrix denotes a popularity vector of a user which reflects the preference of a user for a certain content in a statistical way. Assuming that the content popularity and user preference are stochastic, we can obtain the relation:

$$\mathcal{P}_{uf} = \sum_{u=1}^{N} w_u q_{uf}, \tag{2}$$

where w_u is the probability of user $u \in \mathcal{U}$ sents a request for various contents $f \in \mathcal{F}$, given to a user request distribution $\mathcal{W} = [w_1, w_2, ..., w_N]$, $\sum_{u=1}^{N} w_u = 1$, $w_u \in [0,1]$, which reflects the request active level of each user.

B. D2D Sharing Model

Under the D2D-aid cellular networks, users can select both D2D links model and cellular links model. In the D2D links model, users can request and receive the content from the others via D2D links (e.g., Wi-Fi or Bluetooth) or the users can request the content from the BSs directly in a cellular links manner. In our model, the users select D2D links model in prior. If the requested content is not in their own buffers (or their neighbours'), the cellular links model is chosen.

To model the D2D sharing activities among mobile users, the opportunistic encounter (e.g., user mobility, meeting probability and geographical distance) and social relationship (e.g., online relations and user preference) are two important factors which should be concerned about.

1) $Opportunistic\ Encounter$: In the real world, the users can connect with each other in the D2D links manner. the distance of two users is smaller than a critical value d_c. Since the devices are carried by humans or vehicles, we use the meeting probability to describe the user mobility.

Similar with the prior work [29], we regard λ_{uv} as the contact rate of user u and v which follows the Poisson distribution and the contact event is independent of the user preference. We can obtain the opportunistic delivery as the Poisson process with rate $\mathcal{P}_{uf}\lambda_{uv}$. If user u caches content f in its buffer, we can derive the probability p_{uv} that user v receiving content f from user u before the content expires at time T_f. For a node pair, we can derive that:

$$p_{uv} = \int_{T_f}^{\infty} \mathcal{P}_{uf}\lambda_{uv} e^{-\mathcal{P}_{uf}\lambda_{uv} y} dy = 1 - e^{-\mathcal{P}_{uf}\lambda_{uv} T_f} \tag{3}$$

However, if the content f does not cached in user u, $p_{uv} = 0$. Combined with the definition of s_{uf}, we can overwrite the Equation 3, as $p_{uv} = 1 - e^{-\mathcal{P}_{uf}\lambda_{uf} T_f s_{uf}}$. Hence, the probability that user v cannot receive content f from all the other user $u \in \mathcal{U}$ is $\prod\limits_{u \in \mathcal{U}} (1 - p_{uv})$. Then the probability of user v receive content f from user u can be expressed by

$$P_{uv} = 1 - \prod_{u \in \mathcal{U}} (1 - p_{uv}) = 1 - e^{-\mathcal{P}_{uf} T_f \sum\limits_{u \in \mathcal{U}} \lambda_{uf} s_{uf}}. \tag{4}$$

2) $Social\ Relationship$: In social relationship among users, mobile users with weak social relationship may not be willing to share the content with the others owing to the security/privacy concerns. On the other hand, users sometimes have additional resource and are willing to share the content with others. However, the sharing activities may fail because of the hardware/bandwdth restriction (the content may be too large or the traffic speed is too slow). Thus, we consider the social relationship mainly depends on user preference and content transmission rate condition.

We employ the notion of Cosine Similarity to measure the user preference between two users and the preference similarity factor C_{uv} is defined as

$$C_{uv} = \frac{\sum_{f \in \mathcal{F}} q_{uf} q_{vf}}{\sqrt{\sum_{f \in \mathcal{F}} (q_{uf})^2} \sqrt{\sum_{f \in \mathcal{F}} (q_{vf})^2}}, \forall u, v \in \mathcal{U} \tag{5}$$

Finally, based on the opportunistic encounter and social relationship, we can obtain the probability of D2D sharing between user u and v as follow:

$$P_{uv}^{D2D} = C_{uv} \cdot P_{uv}, \forall u, v \in \mathcal{U}, \forall f \in \mathcal{F}, \tag{6}$$

where $\sum_{v \in U} P_{uv}^{D2D} \leq 1, \forall u \in U$. The sum of probability of D2D sharing between each user and other users is less than 1.

C. Association of Users and BSs

Users can ask the content directly from the associated local BS when the requested content cannot be satisfied by D2D sharing. Definition P_u^{BS} is the cellular serving ratio, which is the average probability that the requests of user u have to be served by local BS via backhual link rather than D2D commnications. Thus, we can obtain that $P_u^{BS} = 1 - \sum_{v \in U} P_{uv}^{D2D}, \forall u \in U$. In this paper, we consider that the content transmission process can be finished within the user mobility tolerant time, e.g. Before the user moves out of the communication range of the local BS. The requested content can be satisfied from the buffer of local BS or obtained from the neighbour BSs via BS-BS link as well as the Internet via backhual link. Let P_{uB}^{BS} denotes the probability of BS n servering user u, we have:

$$P_{uB}^{BS} = \frac{\sum_i T_{un}^{BS}(i)}{\sum_{n \in \mathcal{N}} \sum_i T_{un}^{BS}(i)}, \tag{7}$$

where $T_{un}^{BS}(i)$ denote the time period of the i-th cellular serving from BS n to user u during the total sample time T_{tot}. Therefore, we have the probability P_{uv}^{BS} that user u is served by BS n as follow:

$$P^{BS}_{uB_n} = P^{BS}_{u} \cdot P^{BS}_{uB}, \forall u \in \mathcal{U}, \forall n \in \mathcal{N} \tag{8}$$

Note that $\sum_{n\in\mathcal{N}} P^{BS}_{un} + \sum_{u\in\mathcal{U}} P^{D2D}_{uv} = 1$, $\forall u \in \mathcal{U}$.

D. Caching Placement and Content Delivery

Considering that the content offload at BS and share among users are dynamic, based on the measurement of **Xender** real trace, shown as Fig. 3. We can divide each period into **peak time** and **off-peak time**. During each period, there are two phases, i.e. *caching placement phase and content delivery phase*.

1) Caching Placement: In the caching placement phase, the contents are placed following the Zipf distribution both in the users' devices and BSs [30] uniformly initially. The BS first checks the cache state matrix Φ and the request distribution $\mathcal{W}$, then makes the optimal caching policy by an adaptive learning method. According to the caching replacement policy, the contents are pushed to all the nodes at **off-peak time** by the local BS.

2) Content Delivery: In the content delivery phase, the contents are requested by the individual user based on its preference. A user will ask the others for the requested content if the content is not stored in the user's local cache. Firstly, the user can fetch the requested content from its on-hop neighbour in the D2D link manner within distance smaller than the critical value d_c. This process is called D2D sharing. Otherwise, it will access the local BS directly to request the content.

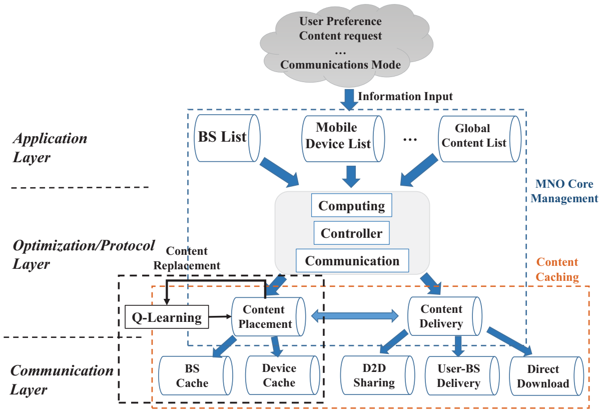


Fig. 2. Brief process of content-centric control and management in D2D aided mobile networks.

E. Content-centric Control and Management

Fig. 2 shows the brief process of content-centric control and management in D2D enabled wireless mobile networks. We next introduce the MNO core, BSs, and mobile devices based on the perspective of content caching, respectively.

1) MNO Core: The computing, services controller and communication resources are maintained in the MNO core. It is a central management to make decisions on content placement and delivery. Global information in the network are organized in MNO core, including BS list, mobile device list and global content list.

- BS List: In the BS list, all the associated BSs are recorded in MNO core, including BS ID, content ID in each BS, mobile device ID connected with each BS, etc. Particularly, in this paper, the content popularity matrix $\mathcal{P}$, cache state matrix Φ of mobile users and user request distribution $\mathcal{W}$ are also maintained in the BS list.
- Mobile Device List: It maintains the information of mobile device ID, historical information of associated BS ID and mobile device ID, association time, geographic mobility pattern, etc.
- Global Content List: It stores all the information of cached or incoming content in the BSs and mobile devices. Particularly, the basic characteristics of contents in the network are also maintained in this list, such as content ID, content size, content popularity, mobile users' preference, ect.

2) BSs: In this paper, BSs act as the role of local management of all the associated mobile devices including the information of mobile device list that registered in the BS, such as their D2D sharing, and provide services to mobile users via cellular links. The computing process of Q-learning for the content replacement is also conducted in the local BS.

3) Mobile Devices: The capacity of caching, computing and communication of each mobile device are limited. A list of the social relationships between its owner and the encountered mobile users, the duration associated time, meeting time, content size/ID and so on, are also maintained in the mobile devices caches.

4) Communications: There are four kinds of communications in the network, the D2D communication, the BS-BS communication, cellular communication and backhaul communication. The mobile devices and BSs can communicate with each other in the way of D2D links and cellular links. The collaboration of contents delivery between BS-BS or BS-SP via the MNO core are provided through BS-BS links and backhaul links, respectively.

IV. Caching Policy Optimization

In the hierarchical wireless networks with cache-enabled D2D communications, we explore the maximum capacity of the network based on the mobility and social behaviors of users. The goal is to optimize the network edge caching by offloading the contents to users via D2D communications and reducing the system cost of content exchange between BSs and Core Network via cellular links.

A. Optimization for D2D Caching Problem

Mobile users can share the content via D2D communications. For user pair u and v, v can get the requested content f from u if u has the content (e.g., $s_{uf} = 1$) when v does not (e.g., $s_{vf} = 0$) with the probability of P^{D2D}_{uv}. Thus, the content offload from the BSs or Internet via D2D link between

u and v can be obtained as $l_f P_{uv}^{D2D}$, l_f is the size of content f. Whether the user u has the content f or not, we can obtain the total content O_{D2D} via D2D sharing as:

$$O_{D2D} = \sum_{f \in F} l_f \sum_{u \in U} P_{uv}^{D2D} s_{uf} (1 - s_{vf}) \tag{9}$$

Our aiming is to maximize the total size of content offload at users via D2D sharing while satisfying all the buffer size constraints of mobile users. Formally, the optimization problem is defined as

$$\begin{aligned} &\max O_{D2D} \\ &s.t. \sum_{f \in \mathcal{F}} s_{uf} l_f \leq L_u, \forall u \in \mathcal{U} \\ &\quad s_{uf} \in \{0,1\}, \forall u \in \mathcal{U}, \forall f \in \mathcal{F}, \end{aligned} \tag{10}$$

where $\sum_{f \in F} s_{uf} l_f \leq L_u$ is the buffer size constraint of all the mobile users' devices and $s_{uf} \in \{0,1\}$ is the cache state in each mobile device.

The optimization problem (10) is NP-hard.

Proof: Let $z_{uv,f} = s_{uf}(1 - s_{vf})$, and $z_{uv,f} \in \{0,1\}$. Thus, we can rewrite the problem (10) as:

$$\begin{aligned} &\max \sum_{f \in \mathcal{F}} l_f \sum_{u \in \mathcal{U}} P_{uv}^{D2D} s_{uf} (1 - s_{vf}) \\ &s.t. \sum_{f \in \mathcal{F}} s_{uf} l_f \leq L_u, \forall u \in \mathcal{U} \\ &\quad z_{vu,f}, s_{uf} \in \{0,1\}, \forall u, v \in \mathcal{U}, \forall f \in \mathcal{F}, \end{aligned} \tag{11}$$

where $\sum_{f \in F} s_{uf} l_f \leq L_u$ is the cardinality constant of L_u. It is easy to observe that the Problem (10) has the same structure with the problem formulated in [31], which has been proved as NP-hard. ■

B. System Cost Analysis

We introduce a suboptimal problem to approximately solve the challenging optimization problem 10. The basic idea is that the whole information (e.g., cache state matrix, user preference, user request distribution, etc.) is maintained in each local BS, we use Markov Decision Process (MDP) to model the cache replacement problem and obtain the best caching policy by learning the user preference and predict the user request distribution over the minimum system cost.

If the requested content has to be served via cellular link between the BSs and Core network, we regard the cellular traffic as the system cost. To ensure the requested content is in BS, let s_{uBf} denote the content f cache state in the BS and $s_{uBf} = s_{Bf}(1 - s_{uf})$. We can obtain the system cost:

$$C = \sum_{n \in N} P_{uB_n}^{BS} (1 - s_{uBf}), \tag{12}$$

where $1 - s_{uBf}$ indicates that there are no requested contents cached in the local BS and the neighbor BSs.

V. Q-Learning Based Cache Replacement

In this section, we model the cache replacement problem as a MDP and propose a distributed Q-learning based cache replacement strategy.

S : cache state in Φ, $S = (s_1, s_2, ..., s_n)$, $n \in \mathcal{N}$, which denotes the set of caching states at all small-cell nodes, $s_n = (s_n(f_1), s_n(f_2), ..., s_n(f_M))$, where $s_n(f) = 1$ represents content f is cached in user n, 0 is the otherwise;

A : replacement action, $A = (a_1, a_2, ...a_n)$, where $a_n = \{a_n(f_m^+, f_k^-) | n \in \mathcal{N}, f_m, f_k \in \mathcal{F}\}$ denotes that the local BS replaces f_k^- with f_m^+ from all nodes under the communication range at off-peak time;

$R(s,a)$: the reward function that when BS determines the action a at state s (i.e. the expected change of system cost caused by the cache replacement policy).

During the MDP of caching replacement, the process of BS n at time t is demonstrated as follows:

- The BS n checks the current cache state in matrix Φ and obtains the content user preference $\mathcal{P}_{uf}$ and the user request distribution $\mathcal{W}$ under current cache state s_n^t.
- Based on the current cache state s_n^t BS n takes the corresponding replacement action a_n^t.
- The environment is transformed into the next state s_n^{t+1} based on the replacement action a_n^t and the reward $R_n^t = R(s_n^t, a_n^t)$ can be obtained due to the transition.
- The reward is fed back to the BS and the process is repeated.

Define policy $\pi : S \rightarrow A$, which guides the next move $a = \pi(s,a)$ by given the current state s, and each policy π is corresponding to a value function:

$$V_\pi(s,a) = E\left[R(s_1, a_1) + \gamma_2 R(s_2, a_2) + ... | s_1 = s, \pi\right] \tag{13}$$

Given a fixed strategy, the value function satisfies the **Bellman Equation**:

$$V_\pi(s,a) = R(s,a) + \gamma \sum_{s' \in S} V_\pi(s', a') \tag{14}$$

We can find the optimal strategy π under the current cache state s by solving the V_π, and define the optimal value function $V^*(s,a)$ as follows:

$$V^*(s,a) = \max_{a \in A} V_\pi(s,a) \tag{15}$$

According to Eq. 13, we can obtain that:

$$V^* = \sum_{t=0}^{\infty} \gamma^t R^t(s,a) \tag{16}$$

Let $\pi^*(s)$ denotes the solution of MDP problem, which provides the optimal policy for action A to give the best move at next time. The optimal policy $\pi^*(s)$ can be obtained by maximizing the reward over an infinite time.

$$V^*(s,a) = R(s,a) + \max_{a \in A} \gamma \sum_{s' \in S} V^*(s', a'), \tag{17}$$

where $0 \leq \gamma \leq 1$ is the discount factor that determines the effect of future rewards to the current decisions.

We apply the Q-learning to find the optimal policy $\pi^*(s,a)$ that corresponds to estimate the V^*. The relationship between V^* and the Q-value can be expressed as

$$V_n^*(s,a) = \max_{a \in A} Q_n^*(s,a) \tag{18}$$

The optimal policy can be determined by $\pi^*(s) = \arg\max_{a \in A} Q^*(s,a)$. Therefore, we can obtain the optimal Q^* by the Q-learning algorithm in a recursive manner for BS n.

$$\begin{aligned} & Q_n^{t+1}(s_n, a_n) \\ & = (1-\alpha) Q_n^t(s_n, a_n) + \alpha \left(R_n^t(s_n, a_n) + \gamma V_n^t(s_n + a_n)\right) \\ & = Q_n^t(s_n, a_n) + \alpha(R_n^t(s_n, a_n) + \gamma V_n^t(s_n + a_n) \\ & -Q_n^t(s_n, a_n)), \end{aligned} \tag{19}$$

where α is the learning rate, $s_n + a_n$ is the transfer state from state s_n after taking the action a_n at time t, and $V_n^t(s_n + a_n) = \max_{a_n \in A} Q_n^t(s,a)$.

Reward function is the most important part of Q-learning algorithm, and it directly decides which content should be replaced. In this paper, our aim is to maximize the content offloading in the D2D sharing manner and satisfy the users' requests locally. Thus, the traffic of cellular serving can be regarded as the system cost, the key idea of designing the reward function is to achieve the highest score when a content is replaced. Formally, we define the reward function as

$$R^t(s_n, a_n) = R^t(s_n(f_m^+), a_n) - R^t(s_n(f_k^-), a_n), \tag{20}$$

where $R^t(s_n, a_n)$ is the reward when performing the action a_n at state s_n, $R^t(s_n(f_m^+), a_n)$ is the reward gain assigned for action a_n by adding the candidate content f_m, denoted as f_m^+. And $R^t(s_n(f_k^-), a_n)$ represents the reward loss when removing the content f_k from the system, denoted as f_k^-. The reward function is computed as the gain minus the loss. According to the Eq. 12, firstly the reward gain can be calculated by the system cost when adding the content f_m as follows:

$$R^t(s_n(f_m^+), a_n) = l_{f_m^+} C_{f_m^+} = l_{f_m^+} \sum_{n \in N} P_{uB_n}^{BS} (1 - s_{uBl_{f_m^+}}), \tag{21}$$

where $l_{f_m^+}$ is the size of candidate content fp_m. Similarly, the reward loss can be expressed as

$$R^t(s_n(f_k^-), a_n) = l_{f_k^-} C_{f_k^-} = l_{f_k^-} \sum_{n \in N} P_{uB_n}^{BS} (1 - s_{uBl_{f_k^-}}) \tag{22}$$

Finally, the system reward can be expressed as

$$\begin{aligned} & R^t(s_n, a_n) \\ & = R^t(s_n(f_m^+), a_n) - R^t(s_n(f_k^-), a_n) \\ & = l_{f_m^+} \sum_{n \in N} P_{uB_n}^{BS} (1 - s_{uBl_{f_m^+}}) - l_{f_k^-} \sum_{n \in N} P_{uB_n}^{BS} (1 - s_{uBl_{f_k^-}}) \end{aligned} \tag{23}$$

VI. Experiment

In this section, we evaluate the proposed cache policy based on the experimental results of the mobile application Xender.

A. Dataset

Xender is one of the world's main applications for file transfer and sharing. It provides users with transmission between mobile devices of different types and sizes of the convenience of file (such as Android, iOS and Windows) via D2D communications. the transmission is free, and doesn't need cables or Wi-Fi or cellular network connection, and definitely doesn't need a mobile data transmission. Currently *Xender* has around 10 million daily and 100 million monthly active users, as well as about 110 million daily content deliveries.

We capture *Xender*'s trace for one month (from 01/09/2016 to 30/09/2016), including 574,465 active mobile users, conveying 131,588 content files, and 533,146,832 content requests.

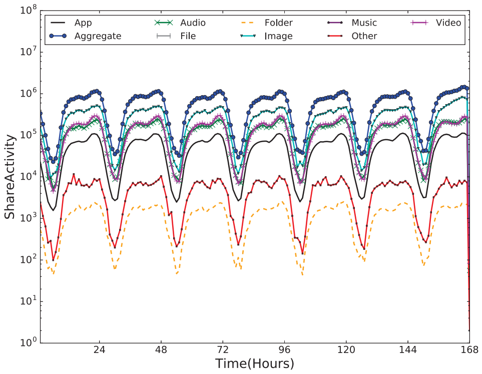
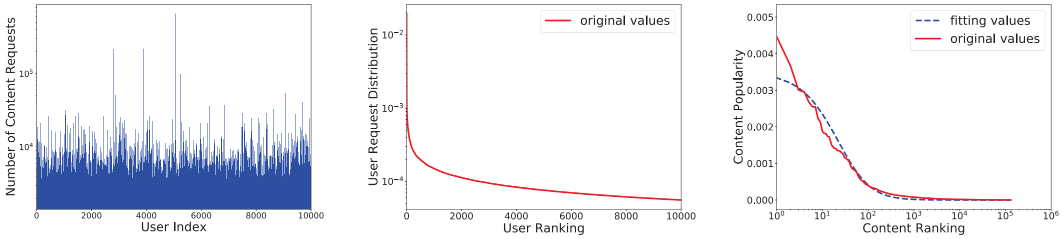


Fig. 3. Statistic for all Sharing Activities

B. Parameter Settings

From Fig.3 [32], we choose 1 week from the dataset to show the number of sharing activities over time with regard to various types' contents and aggregated content value. We can obtain that each week and day shows a typical life cycle with temporal regularity, and further verifies the fact that families and friends always get together in the weekend in India. This implies that in real environment there exists large room of optimizing the time-varying system resource and proposing effective family-focused marketing strategies.

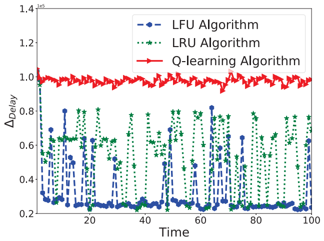
We extract 10,000 mobile users from *Xender*'s trace to demonstrate different numbers of content requests from different users and user request distribution in Fig. 4(a) and Fig.

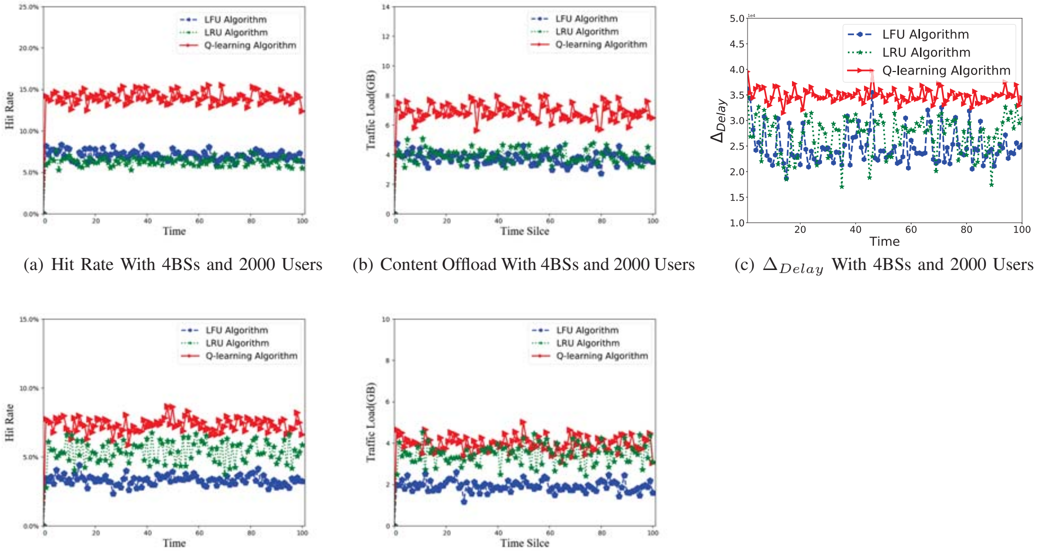


(a) Content requests of users

(b) User Request Distribution

Fig. 4. Distributions of users' content requests, content popularity and user's requ





(a) Hit Rate With 4BSs and 2000 Users

(b) Content Offload With 4BSs and 2000 Users

(c) Δ_{Delay} With 4BSs and 2000 Users

(d) Hit Rate With 8 BSs and 1000 Users through Q-learning

(e) Content Offload With 8 BSs and 1000 Users through Q-learning

(f) Δ_{Delay} With 8 BSs and 1000 Users through Q-learning

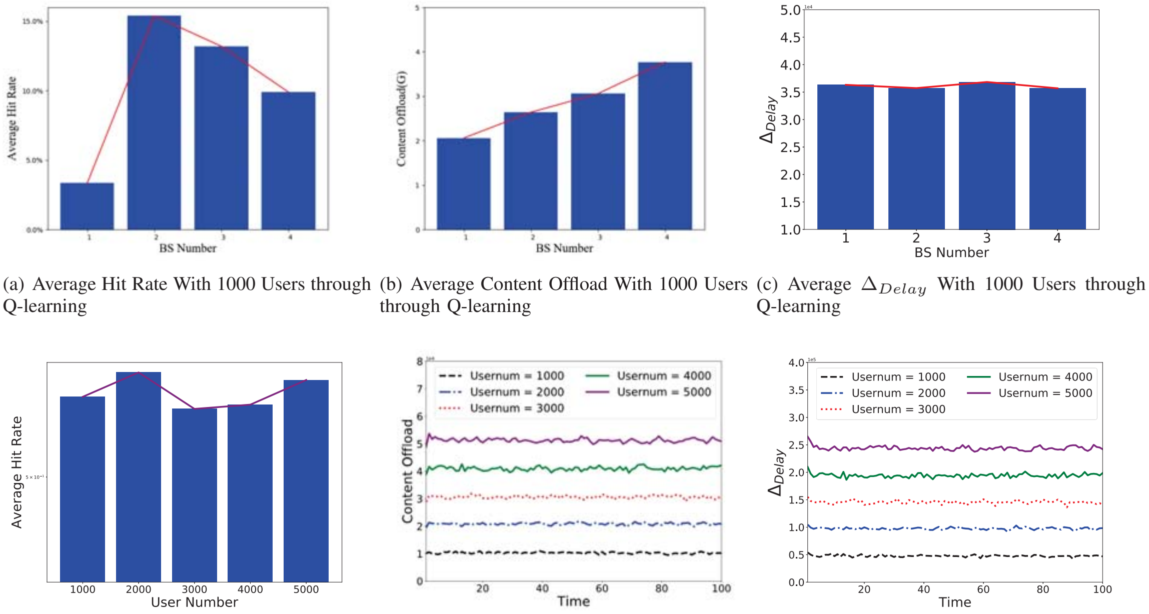
Fig. 5. Performance Demontration of Hit Rate, Content Offload and Delay With LRU, LFU and Q-learning

4(b) . As shown in Fig. 4(c), content popularity distribution can be well fitted by a MZipf distribution with τ of 1.58. The above results are based on Xender real data set analysis.

C. Cache Simulation Results and Discussions

In the simulation platform, we evaluate the performance of the strategy through 3 metrics the hit rate, content offload and Δ_{Delay}.

- **Hit Rate**: where N_h is the number of satisfied user requests, N_r is the total number of request. From the second iteration of the download situation, there are many requests for a time slice, while other requests are new requests, except for the required download status.
- **Content Offload**: Each content download has a fixed time slice limit, for example, there is a file size of 30, which needs 3 time slices to download, and we think that each time slice can download one third of the content, so the content offload of a time slice is 10.
- Δ_{Delay}: In Tab. I, we set the time delay of the user to the base station to be 10ms, the time delay of the base station to the base station is 5ms, the time delay of the base station to MNO is 20ms, and the time delay of MNO to the Internet is 100ms. In Eq. (25) and Eq. (26), we use Δ_{Delay} as the criterion, and Δ_{Delay} is calculated according to different scenarios.



(a) Average Hit Rate With 1000 Users through Q-learning

(b) Average Content Offload With 1000 Users through Q-learning

(c) Average Δ_{Delay} With 1000 Users through Q-learning

(d) Hit Rate With 4 BSs through Q-learning

(e) Content Offload With 4 BSs through Q-learning

(f) Δ_{Delay} With 4 BSs through Q-learning

Fig. 6. Performance Demontration of Hit Rate, Content Offload and Delay Under Different Number of BSs and Users Of Q-learning

If the direct hit occurs, Eq. (25) is workable, where D^{B}_{MNO} is the value of delay from BS to MNO, D^{MNO}_{Int} is from MNO to Internet and N_t is the number of time slices.

$$\Delta_{Delay2} = \frac{D^{B}_{MNO} + D^{MNO}_{Int} - D^{B}_{B}}{N_t} \qquad (26)$$

We use Eq. (26) when hit is from the associated base station, D^{B}_{B} is the value of delay from BS to BS, other parameters are same with them in Eq. (26).

TABLE I
EXPERIMENT SETTING

User to BS Transmission Delay	10 ms
User to User Transmission Delay	5 ms
BS to MNO Transmission Delay	20 ms
MNO to Internet Transmission Delay	100 ms
User Numbers	1000 to 5000
BSs Numbers	1 to 4, 8
Content Numbers	120

Compared with LRU and LFU, we demonstrate the different experimental evaluations of the three metrics in Fig. 5 and Fig. 6. Specifically, we compared the hit rate and content offload with these 3 strategies. The value of hit rate and content offload in Q-learning is better than others. For example, in Fig. 5(a), Fig. 5(b) and Fig. 5(c), we set the number of base stations in the region as 4, the number of mobile users is 2000, and the content file number is 120. We can observe that the proposed caching strategy achieves the best performance. For example, the hit rate of Q-learning based strategy is about 15% while the values of the other two strategies are around 8% in Fig.5(a). The result of content offload in Q-learning can increase by more than 50% over LRU and LFU. Compared with LRU and LFU, the delay by Q-learning is reduced 20%. In Fig. 5(e) and Fig. 5(f), When the number of base station is 3 or 4, the effect is better.

To better explore the Q-learning Strategy. We compare the performance of Q-learning proposed under different number of base stations and different number of mobile users. It is shown in Fig. 6(a) Fig. 6(b) and Fig. 6(c), as the number of base stations increases, in 1000 mobile users region, the values of traffic and Δ_{Delay} are rises, but the value of hit rate drops. When the number of base station is 4, the content offload which value is 3,696 MB is the highest. When the number of base station is 3, the value of content offload is 3,153 MB and the Δ_{Delay} is 3,6841 ms.

Besides, we also change the number of users to explore the caching policy performance. The number of mobile users is set from 1000 to 5000. It is shown in the Fig. 6(d) , with the increase of number of users, the value of hit rate tends to be stable, when user number is 2000, it is the highest increasement with 53.4%. Fig. 6(e) and Fig. 6(f) show the content offloading and content access delay, the performance has been steadily increasing.

In the experiment, we compared the effects of LRU, LFU and Q-learning by hit rate, content offload and Δ_{Delay}. The values of hit rate are 27%, 18%, 53%, and Q-learning is obviously better than LRU and LFU. The result of content offload in Q-learning can increase by more than 38% over LRU and 70% over LFU. Compared with LRU and LFU, the delay by Q-learning is reduced more than 20%. And we also demontrate the performance of hit rate, content offload and Δ_{Delay} under different number of BSs and users Of Q-learning. When the number of base station is 4, the content offload which value is 8,596 MB is the highest. And when user number is 2000, the value of hit rate is the highest increasement with 53.4%. From these figures, we can see the performance of the strategy we proposed is the best.

VII. Conclusions

In this paper, we have proposed an edge caching policy of hierarchical wireless networks by system learning and analysis from the user mobility and social relationship, aiming to maximize the size of content offload via D2D sharing. Specifically, we have employed with Markov Decision Process and a Q-learning in the proposed content replacement strategy to replace the contents of all the users at **off-peak time**. To realize the proposed strategy, we have analyzed the system cost to calculate the reward and achieve the best replacement policy. The results of real trace based experiment have shown that our proposed strategy can always outperform the LRU and LFU schemes in terms of request hit rate, traffic offload and the saved delay.

Acknowledgment

This work is supported by the National Natural Science Foundation of China through Grant No. 61702364.

References

[1] H. Zhou, H. Wang, X. Li, and V. C. M. Leung, "A Survey on Mobile Data Offloading Technologies," IEEE Access, vol. 6, pp. 5101-5111, Jan. 2018.

[2] X. Wang, M. Chen, Z. Han, et al. "TOSS: Traffic offloading by social network service-based opportunistic sharing in mobile social networks" in Proc. INFOCOM, 2014, pp. 2346-2354.

[3] M. Gregori, J. G. Vilardebo, J. Matamoros, and D. Gunduz, "Wireless Content Caching for Small Cell and D2D Networks," IEEE J. Sel. Areas Commun., vol. 34, no. 5, pp. 1222-1234, May 2016.

[4] D. Feng, L. Lu, Y. Y. Wu, G. Y. Li, S. Li, and G. Feng, "Device-to-Device Communications in Cellular Networks," IEEE Commun. Mag., vol. 52, no. 4, pp. 49-55, Apr. 2014.

[5] J. Tang, G. Xue and C. Chandler, "Interference-aware routing and bandwidth allocation for QoS provisioning in multihop wireless networks", Wireless Communications and Mobile Computing (WCMC), Vol. 5, No. 8, 2005, pp. 933-944.

[6] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang and J. Wang, "Geospatial and temporal dynamics of application usage in cellular data networks", IEEE Transactions Mobile Computing, Vol. 14, No. 7, 2015, pp. 1369-1381.

[7] M. P. Wittie, V. Pejovic, L. Deek, K. C. Almeroth, and B. Y. Zhao, "Exploiting Locality of Interest in Online SNS," ACM CoNEXT, 2010.

[8] N. Morozs, T. Clarke, et al. "Distributed Heuristically Accelerated Q-Learning for Robust Cognitive Spectrum Management in LTE Cellular Systems". IEEE Transactions on Mobile Computing, 2016, 15(4):817-825.

[9] X. Wang, Y. Zhang, V. C. M. Leung, et al. "D2D Big Data: Content Deliveries over Wireless Device-to-Device Sharing in Large-Scale Mobile Networks". IEEE Wireless Communications, vol. 25, no. 1, pp.32-38, February, 2018.

[10] B. N. Bharath, K. G. Nagananda, H. V. Poor. "A Learning-Based Approach to Caching in Heterogenous Small Cell Networks". IEEE Transactions on Communications, 2016, 64(4):1674-1686.

[11] M. Srinivasan, V. J. Kotagi, C. S. R. Murthy. "A Q-Learning Framework for User QoE Enhanced Self-Organizing Spectrally Efficient Network Using a Novel Inter-Operator Proximal Spectrum Sharing". IEEE Journal on Selected Areas in Communications, 2016, 34(11):2887-2901.

[12] Y. Zhang, B. Song, P. Zhang. "Social behavior study under pervasive social networking based on decentralized deep reinforcement learning". Journal of Network and Computer Applications, 2016, 86.

[13] T. Rodrigues, F. Benvenuto, M. Cha, K. Gummadi, and V. Almeida, "On Word-of-Mouth Based Discovery of the Web," ACM IMC, 2011.

[14] 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 Commun. Mag., vol. 52, no. 2, pp. 131-139,Feb. 2014.

[15] E. Zeydan et al., "Big data caching for networking: Moving from cloud to edge," IEEE Commun. Mag., vol. 54, no. 9, pp. 3642, Sep. 2016.

[16] 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 Commun. Mag., vol. 51, no. 4, pp. 142-149, Apr. 2013.

[17] 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, Mar. 2012, pp. 1107-1115.

[18] B. Han, X. Wang, N. Choi, T. K. Kwon, and Y. Choi, "AMVS-NDN: Adaptive mobile video streaming and sharing in wire less named data networking," in Proc. IEEE INFOCOM Workshop, Apr. 2013.

[19] X. Li, X. Wang, S. Xiao, and V. C. M. Leung, "Delay performance analysis of cooperative cell caching in future mobile networks," in Proc. IEEE ICC, Jun. 2015, pp. 5652-5657.

[20] J.-P. Hong and W. Choi, "User prefix caching for average playback delay reduction in wireless video streaming," IEEE Trans. Wireless Commun., vol. 15, no. 1, pp. 377-388, Jan. 2016.

[21] X. Chen, J. Wu, Y. Cai, H. Zhang, and T. Chen, "Energy-efficiency oriented traffic offloading in wireless networks: A brief survey and a learning approach for heterogeneous cellular networks," IEEE J. Sel. Areas Commun., vol. 33, no. 4, pp. 627-640, Apr. 2015.

[22] D. Liu and C. Yang, "Energy efficiency of downlink networks with caching at base stations," IEEE J. Sel. Areas Commun., vol. 34, no. 4, pp. 907-922, Apr. 2016.

[23] A. Liu and V. K. N. Lau, "Exploiting base station caching in MIMO cellular networks: Opportunistic cooperation for video streaming," IEEE Trans. Signal Process., vol. 63, no. 1, pp. 57-69, Jan. 2015.

[24] X. Li, X. Wang, K. Li, et al. "Collaborative Multi-tier Caching in Heterogeneous Networks: Modeling, Analysis, and Design," IEEE Trans. on Wireless Commun., vol. 1, no. 1, PP. 99, Agust, 2017.

[25] K. Poularakis and L. Tassiulas, "On the complexity of optimal content placement in hierarchical caching networks," IEEE Trans. Commun., vol. 64, no. 5, pp. 2092-2103, May 2016.

[26] J. Dai, Z. Hu, B. Li, J. Liu, and B. Li, "Collaborative hierarchical caching with dynamic request routing for massive content distribution," in Proc. IEEE INFOCOM, Mar. 2012, pp. 2444-2452.

[27] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks," IEEE Commun. Mag., vol. 52, no. 8, pp. 82-89, Aug. 2014.

[28] M. Hefeeda and O. Saleh, "Traffic Modeling and Proportional Partial Caching for Peer-to-Peer Systems," IEEE/ACM Trans. Netw., vol. 16, no. 6, pp. 1447-1460, Dec. 2008.

[29] A. Balasubramanian, B. Levine, and A. Venkataramani, "DTN Routing as a Resource Allocation Problem," Proc. ACM SIGCOMM Conf. Applications, Technologies, Architectures, and Protocols for Computer Comm., pp. 373-384, 2007.

[30] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das, "Understanding traffic dynamics in cellular data networks," in Proc. IEEE INFOCOM, 2011.

[31] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, "Introduction to Algorithms", 2nd ed. Cambridge, MA, USA: MIT Press, 2001.

[32] Wang, Shanjia, et al. "Large scale measurement and analytics on social groups of device-to-device sharing in mobile social networks." Mobile Networks and Applications (2017): 1-13.