Service Delay Minimization-Based Joint Clustering and Content Placement Algorithm for Cellular D2D Communication Systems

Ahmad Zubair, Pengfei Ma, Tao Wei, Ling Wang, and Rong Chai(✉)

Key Lab of Mobile Communication Technology,
Chongqing University of Posts and Telecommunications, Chongqing 400065, China
chairong@cqupt.edu.cn

Abstract. The rapidly increasing content fetching requirements pose challenges to the transmission performance of traditional cellular system. Due to the limited transmission performance of cellular links and the caching capabilities of the base stations (BSs), it is highly difficult to achieve the quality of service (QoS) requirements of multi-user content requests. In this paper, a joint user association and content placement algorithm is proposed for cellular device-to-device (D2D) communication network. Assuming that multiple users located in a specific area may have content requests for the same content, a clustering and content placement mechanism is presented in order to achieve efficient content acquisition. A joint clustering and content placement optimization model is formulated to minimize total user service delay, which can be solved by Lagrange partial relaxation, iterative algorithm and Kuhn-Munkres algorithm, and the joint clustering and content placement strategies can be obtained. Finally, the effectiveness of the proposed algorithm is verified by MATLAB simulation.

Keywords: Cellular network · Device-to-device D2D communication · User association · Content placement · Service delay

1 Introduction

The rapid proliferation of new applications poses great challenges to the traditional cellular systems. To improve user quality of service (QoS) as well as network performance, device-to-device (D2D) communication technology can be applied in cellular systems which allows adjacent user equipments (UEs) communicate with each other in a direct manner without relying on the data forwarding of the base stations (BSs) [1]. Benefited from the improved channel characteristics between D2D peers, D2D communication technology is expected to improve system throughput, reduce transmission delay and power consumption of the devices significantly.

Transmission mode selection and resource allocation problem in cellular D2D communication systems was addressed in previous research work [5–8]. In [6], the authors considered the transmission mode selection problem and presented an energy consumption minimization-based optimal scheme. In [7], the joint transmission mode selection

H. Gao et al. (Eds.): ChinaCom 2019, LNICST 313, pp. 180–191, 2020.
https://doi.org/10.1007/978-3-030-41117-6_15

and resource allocation problem was formulated as end-to-end sum-rate maximization problem and solved based on BS scheduling method. Considering the constraint on transmission rate, [5] presented a joint transmission mode selection and power control scheme to maximize the energy efficiency of the system. While resource allocation issue was stressed in [5–7], they failed to consider the efficient utilization of resources in the system. Under the assumption of limited resources, the authors in [8] modeled the transmission model selection problem as resource utilization maximization problem, and proposed a channel state-based model selection mechanism to achieve higher resource utilization and D2D transmission gain.

By caching popular contents at the BSs of the cellular system or at certain UEs, the performance of content fetching can be improved significantly. Considering a cellular D2D communication system, the authors in [9] proposed an information-oriented new network architecture enabling wireless network virtualization and D2D communications in order to achieve the maximum revenue of mobile operators. Taking into account various user preference, [10] presented an optimal content delivery strategy which achieves the maximum gain of network offloading.

While user association and content placement have been studied in previous work, the joint design of the two strategies failed to be discussed extensively, thus may result in undesired content fetching performance. Furthermore, while throughput or network revenue were mainly considered in previous work, service delay, which is of particular importance for delay-sensitive users, was not considered for designing joint user association and content placement strategy. In this paper, we consider various content fetching requirements of the users and the content delivery performance in different transmission modes, introducing clustering scheme, and propose a service delay minimization-based joint clustering and content placement algorithm for cellular D2D communication systems.

2 System Model and Proposed Clustering Mechanism

This paper considers a cellular D2D communication system, which is composed of one BS and M content request users (RUs) and M serving users (SUs). Suppose RUs request contents with a certain probability, and the total number of content files required from RUs is denoted by K. Let $P_{i,k}$ denote the preference probability of the ith RU, denoted as RU_i for content k, we obtain $P_{i,k} \in [0,1],\ 0 \leq \sum_{k=1}^{K} P_{i,k} = 1, 1 \leq i \leq M$. The size of content k is denoted by C_k.

In order to achieve efficient content request and reduce content fetching delay, we assume that some popular contents can be cached at the BS or certain SUs, and RUs are allowed to fetch content file in cellular communication mode or D2D transmission mode. More specifically, in cellular communication mode, the RUs access the BS of cellular system to acquire their required content files, while in D2D transmission mode, the RUs interact with their D2D peers, i.e., SUs, to fetch their content files.

For the sake of simplicity, it is assumed that orthogonal frequency division multiple access (OFDMA) scheme is applied for the information interaction in cellular communication mode and D2D transmission mode. As various orthogonal subcarriers are

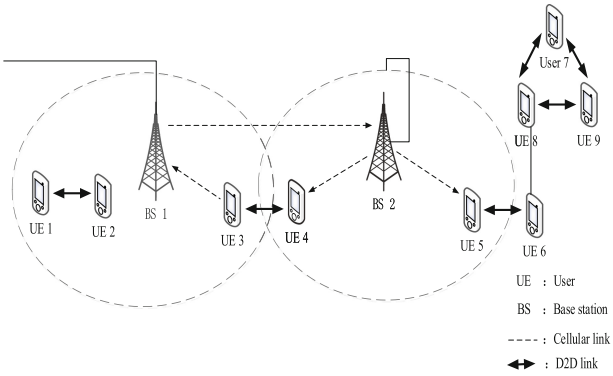

Fig. 1. D2D communication application scenario

allocated for different RUs, there is no transmission interference caused among transmission links. Figure 1 shows the cellular D2D communication system considered in this paper.

Given content file requirements of the RUs and various content fetching performance of the transmission modes, this paper aims to jointly design the transmission mode for the RUs and the content placement strategy for the SUs.

In a cellular D2D communication system, some RUs may need to acquire same content files. By caching some hot contents at certain SUs and employing D2D transmission mode, efficient content access service can be achieved.

Taking into account the diverse content request of RUs and various channel characteristics of the links between users, this paper applies clustering idea and proposes a clustering-based content fetching mechanism. According to network status and user characteristics, the RUs and SUs in the network are dynamically divided into multiple clusters with each cluster consisting of one cluster head (CH) and multiple cluster members (CMs). Without loss of generality, we assume that the CHs are allowed to access the BS directly, while the CMs can only interact with their associated CH.

By suitably choosing SUs as CHs and caching selected content files at the CHs, the CMs may fetch content files through connecting with their CHs in D2D communication mode, In this way, intra-cluster content sharing can be achieved.

Let N_1 denote the maximum number of CHs in the system, and CH_j denote the jth CH, $1 \leq j \leq N_1$. Assuming each CH has a maximum cache capacity for caching content files, we denote $C_j^{\max}$ as the maximum cache capacity of CH_j, $1 \leq j \leq M$. Further assuming that each CH has a limit on the maximum number of associated CMs, we denote N_2 as the maximum number of CMs that associate with one CH.

3 Optimization Problem Formulation

In this paper, we stress the performance of service delay of the RUs, and formulate joint user association and content placement problem as a service delay minimization problem. The detail problem formulation will be discussed in this section.

3.1 User Service Delay Formulation

The total service delay of the users in the system is defined as the sum of intra-cluster D2D communication delay, the delay required for the CHs to fetch contents from the BS, and the content fetching delay in cellular communication mode. We denote the total service delay of the users by D, which can be expressed as

$$D = D^{\mathrm{cm}} + D^{\mathrm{ch}} + D^{\mathrm{b}} \tag{1}$$

where D^{cm} represents the intra-cluster D2D communication delay, D^{ch} denotes the delay required for the CHs to fetch contents from the BS, and D^{b} denotes the delay in cellular communication mode. D^{cm} in (1) can be expressed as

$$D^{\mathrm{cm}} = \sum_{i=1}^{M} \sum_{j=1, j\neq i}^{M} \sum_{k=1}^{K} \delta_{i,j} \beta_{j,k} D^{\mathrm{d}}_{i,j,k} \tag{2}$$

where $\delta_{i,j} \in \{0, 1\}$ is the association variable between RUs and the CHs, i.e., $\delta_{i,j} = 1$, if RU_i associates with CH_j, otherwise, $\delta_{i,j} = 0$; $\beta_{j,k}$ denotes content placement variable, i.e., $\beta_{j,k} = 1$, if content k is placed at CH_j, otherwise, $\beta_{j,k} = 0$; $D^{\mathrm{d}}_{i,j,k}$ denotes the service delay when RU_i associates with CH_j and receives content k, $D^{\mathrm{d}}_{i,j,k}$ can be expressed as

$$D^{\mathrm{d}}_{i,j,k} = \frac{C_k}{R^{\mathrm{d}}_{i,j}} \tag{3}$$

where $R^{\mathrm{d}}_{i,j}$ is the transmission rate of the link between RU_i and CH_j.

In (1), D^{ch} can be calculated as

$$D^{\mathrm{ch}} = \sum_{j=1}^{M} \sum_{k=1}^{K} \delta_{j,j} \beta_{j,k} D^{\mathrm{c}}_{j,k} \tag{4}$$

where $\delta_{j,j}$ indicates that RU_j is selected as a CH, $D^{\mathrm{c}}_{j,k}$ represents the corresponding service delay when the BS sends content k to CH_j, and $D^{\mathrm{c}}_{j,k}$ can be computed as

$$D^{\mathrm{c}}_{j,k} = \frac{C_k}{R^{\mathrm{c}}_{j}} \tag{5}$$

where R^{c}_{j} is the transmission rate of the link between the BS and CH_j.

In (1), D^{b} is given by

$$D^{\mathrm{b}} = \sum_{i=1}^{M} \sum_{k=1}^{K} \left(1 - \sum_{j=1, j\neq i}^{M} \delta_{i,j} \beta_{j,k} \right) P_{i,k} D^{\mathrm{b}}_{i,k} \tag{6}$$

where $D^{\mathrm{b}}_{i,k}$ is the resulted service delay when RU_i associates the BS to obtain content k, $D^{\mathrm{b}}_{i,k}$ can be expressed as

$$D^{\mathrm{b}}_{i,k} = D^{\mathrm{t}}_{i,k} + D^{\mathrm{w}}_{i,k} \tag{7}$$

where $D^{\mathrm{t}}_{i,k}$ and $D^{\mathrm{w}}_{i,k}$ denote respectively the data transmission delay and queueing delay when RU_i associates with the BS and acquires content k.

3.2 Optimization Model

Under the constraints of user clustering, the cache capacity of CHs, and the minimum transmission rate requirements of the RUs, etc, we formulate joint clustering and content placement problem in cellular D2D communication system as a constrained service delay minimization problem, i.e.,

$$\begin{aligned}
&\min_{\delta_{i,j},\beta_{j,k}} \quad D \\
&\text{s.t.} \quad \text{C1}: \delta_{i,j} \in \{0,1\}, \forall i,j \\
&\qquad \text{C2}: \beta_{j,k} \in \{0,1\},\ \forall j,k \\
&\qquad \text{C3}: \sum_{j=1}^{M} \delta_{j,j} \leq N_1 \\
&\qquad \text{C4}: \sum_{i=1,i\neq j}^{M} \delta_{i,j} \leq N_2,\ \forall j \\
&\qquad \text{C5}: \sum_{j=1,j\neq i}^{M} \delta_{i,j} \leq 1,\ \forall i \\
&\qquad \text{C6}: \sum_{k=1}^{K} \beta_{j,k} C_k \leq C_j^{\max},\ \forall j \\
&\qquad \text{C7}: R_i \geq R_i^{\min},\ \forall i
\end{aligned} \tag{8}$$

where C1 and C2 are the binary condition of the CH association variables and the content placement variables, C3 represents the constraint on the maximum number of the CHs, C4 and C5 are respectively the CH association constraint and the CH selection constraint, and C6 is the maximum cache capacity constraint of the CHs. In C7, R_i and $R_i^{\min}$ denote respectively the achievable transmission rate and the minimum transmission rate requirement of RU_i, hence, C7 represents the constraint on the minimum transmission rate requirement of the RUs.

4 Solution to the Optimization Problem

Since the optimization problem given in (8) is a non-convex mixed integer optimization problem, the optimal solution of which is difficult to obtain by the conventional convex optimization algorithm. In this paper, by using the McCormick convex relaxation method [12] and the Lagrangian partial relaxation method [13], the original optimization problem is equivalently converted into three convex optimization subproblems and the modified Kuhn-Munkres (K-M) algorithm [14] is then used to solve the subproblems.

4.1 Reformulation of the Optimization Problem

The optimization problem in (8) contains a number of Boolean variables such as $\delta_{i,j}, \beta_{j,k}$. To tackle the coupling relationship between the variables, we define $\alpha_{i,j,k} = \delta_{i,j}\beta_{j,k}$ and replace $\delta_{i,j}\beta_{j,k}$ by $\alpha_{i,j,k}$ in (8), i.e.,

$$\begin{aligned}\min_{\delta_{i,j},\beta_{j,k},\alpha_{i,j,k}} & \sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{k=1}^{K}\alpha_{i,j,k}D^{\mathrm{c}}_{j,k} + \sum_{i=1}^{M}\sum_{j=1,j\neq i}^{M}\sum_{k=1}^{K}\alpha_{i,j,k}D^{\mathrm{d}}_{i,j,k}P_{i,k} \\ & + \sum_{i=1}^{M}\sum_{k=1}^{K}(1-\sum_{j=1,j\neq i}^{M}\alpha_{i,j,k})P_{i,k}D^{\mathrm{b}}_{i,k} \\ \text{s.t.} \quad & \mathrm{C1}-\mathrm{C7} \\ & \mathrm{C8}: \alpha_{i,j,k} = \delta_{i,j}\beta_{j,k}\end{aligned} \tag{9}$$

C8 in above problem is a non-convex optimization constraint, which can be equivalently converted to the convex optimization constraints C9–C12 by using the McCormick convex relaxation method.

$$\begin{aligned}\min_{\delta_{i,j},\beta_{j,k},\alpha_{i,j,k}} & \sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{k=1}^{K}\alpha_{i,j,k}D^{\mathrm{c}}_{j,k} + \sum_{i=1}^{M}\sum_{j=1,j\neq i}^{M}\sum_{k=1}^{K}\alpha_{i,j,k}D^{\mathrm{d}}_{i,j,k}P_{i,k} \\ & + \sum_{i=1}^{M}\sum_{k=1}^{K}(1-\sum_{j=1,j\neq i}^{M}\alpha_{i,j,k})P_{i,k}, D^{\mathrm{b}}_{i,k} \\ \text{s.t.} \quad & \mathrm{C1}-\mathrm{C7}, \mathrm{C9}-\mathrm{C12} \\ & \mathrm{C9}: \alpha_{i,j,k} \geq 0 \\ & \mathrm{C10}: \alpha_{i,j,k} \geq \delta_{i,j} + \beta_{j,k} - 1 \\ & \mathrm{C11}: \alpha_{i,j,k} \leq \delta_{i,j} \\ & \mathrm{C12}: \alpha_{i,j,k} \leq \beta_{j,k}\end{aligned} \tag{10}$$

To solve the optimization problem in (10), we apply Lagrangian partial relaxation method and relax the constraints C10–C12. In addition, the corresponding Lagrangian multipliers $\eta_{i,j,k},\ \varphi_{i,j,k},\ \theta_{i,j,k}$ are introduced and the non-negative constraints on the Lagrangian multipliers are added in the optimization problem, i.e.,

$$\begin{aligned}\max_{\eta_{i,j,k},\varphi_{i,j,k},\theta_{i,j,k}} \min_{\delta_{i,j,},\beta_{j,k},\alpha_{i,j,k}} & L(\delta_{i,j,},\beta_{j,k},\alpha_{i,j,k},\eta_{i,j,k},\varphi_{i,j,k},\theta_{i,j,k}) \\ \text{s.t.} \quad & \mathrm{C1}-\mathrm{C7},\ \mathrm{C9} \\ & \mathrm{C13}: \eta_{i,j,k} \geq 0 \\ & \mathrm{C14}: \varphi_{i,j,k} \geq 0 \\ & \mathrm{C15}: \theta_{i,j,k} \geq 0\end{aligned} \tag{11}$$

Given the Lagrangian multipliers $\eta_{i,j,k}, \varphi_{i,j,k}, \theta_{i,j,k}$, the Lagrangian function can be expressed as

$$\begin{aligned}& L(\delta_{i,j,},\beta_{j,k},\alpha_{i,j,k},\eta_{i,j,k},\varphi_{i,j,k},\theta_{i,j,k}) \\ & = f_1(\delta_{i,j}) + f_2(\beta_{j,k}) + f_3(\alpha_{i,j,k})\end{aligned} \tag{12}$$

Since there is no coupling between the variables in the three functions $f_1(\delta_{i,j})$, $f_2(\beta_{j,k})$ and $f_3(\alpha_{i,j,k})$, the original dual problem can be converted into three subproblems, i.e., user association subproblem SP1, content placement subproblem SP2, and the joint optimization subproblem SP3.

4.2 Iterative Algorithm-Based Solution

Since the optimization variables and the Lagrangian multipliers in the three subproblems are related, in order to obtain the optimal solution of each subproblem, the optimization variables and Lagrangian multipliers should be solved jointly. To this end, we present an iterative algorithm-based method which calculates the optimization variables and the Lagrangian multipliers successively.

Given the maximum number of CHs N_1, we may consider different CH selection possibilities. Let L denote the number of CH selection strategies. For each particular CH selection strategy, we solve the subproblems respectively based on the given Lagrangian multipliers $\eta_{i,j,k}, \varphi_{i,j,k}, \theta_{i,j,k}$, then compare the obtained total service delay corresponding to various CH selection strategies, and select the joint clustering and content placement strategy which offers the smallest total service delay as the global optimal strategy.

K-M Algorithm-Based Solution to the Subproblems. For the lth CH selection strategy, given the Lagrangian multipliers $\eta_{i,j,k}, \varphi_{i,j,k}, \theta_{i,j,k}$, each subproblem is an integer optimization problem containing binary variables, which can be regarded as the matching problem in a bipartite graph. Hence, for individual subproblem, we may set up the bipartite graph with corresponding vertex set, link set and the weight set of links. Then applying the modified K-M algorithm, we can obtain the user association strategy $\delta_{i,j}^{(l,*)}$, the content placement strategy $\beta_{j,k}^{(l,*)}$, and the joint optimization strategy $\alpha_{i,j,k}^{(l,*)}$.

Lagrangian Multiplier Update. Based on the local optimal solution $\delta_{i,j}^{(l,*)}, \beta_{j,k}^{(l,*)}$, $\alpha_{i,j,k}^{(l,*)}$, the gradient iterative algorithm can be used to update the Lagrangian multipliers. The update formula are:

$$\eta_{i,j,k}(t+1) = [\eta_{i,j,k}(t) - \omega_1(\alpha_{i,j,k}^{(l,*)}(t) + 1 - \delta_{i,j}^{(l,*)}(t) - \beta_{j,k}^{(l,*)}(t)]^+ \tag{13}$$

$$\varphi_{i,j,k}(t+1) = [\varphi_{i,j,k}(t) - \omega_2(\delta_{i,j}^{(l,*)}(t) - \alpha_{i,j,k}^{(l,*)}(t))]^+ \tag{14}$$

$$\theta_{i,j,k}(t+1) = [\theta_{i,j,k}(t) - \omega_3(\beta_{j,k}^{(l,*)}(t) - \alpha_{i,j,k}^{(l,*)}(t))]^+ \tag{15}$$

where $\omega_x, x \in \{1, 2, 3\}$ is the step size.

The algorithm proposed in this paper is shown in Table 1.

Table 1. Proposed joint user association and content placement algorithm

1. Determine L CH selection strategies;
2. Set the maximum number of iterations $T^{\max}$ and the maximum tolerance delay ε;
3. Set l=1;
4. Repeat main program
5. Initialize Lagrangian multipliers $\eta_{i,j,k}, \varphi_{i,j,k}, \theta_{i,j,k}$;
6. Solve user association subproblem, obtain local optimal strategy $\delta'_{i,j}$;
   Solve content placement subproblem, obtain local optimal strategy $\beta'_{j,k}$;
   Solve the joint optimization subproblem, obtain local optimal strategy $\alpha'_{i,j,j}$;
7. Update Lagrangian multipliers
$\eta_{i,j,k}(t+1) = [\eta_{i,j,k}(t) - \omega_1(\alpha'_{i,j,k}(t) + 1 - \delta_{i,j}(t) - \beta'_{j,k}(t)]^+$;
$\varphi_{i,j,k}(t+1) = [\varphi_{i,j,k}(t) - \omega_2(\delta_{i,j}(t) - \alpha'_{i,j,k}(t))]^+$;
$\theta_{i,j,k}(t+1) = [\theta_{i,j,k}(t) - \omega_3(\beta j, k(t) - \alpha'_{i,j,k}(t))]^+$;
8. if $\sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{k=1}^{K}(|\eta_{i,j,k}(t+1) - \eta_{i,j,k}(t)|+$
$|\varphi_{i,j,k}(t+1) - \varphi_{i,j,k}(t)| + |\theta_{i,j,k}(t) - \theta_{i,j,k}(t)|) \leq \varepsilon$
9. algorithm converges,
   return $\delta_{i,j}^{(l),*} = \delta'_{i,j},\ \beta_{j,k}^{(l),*} = \beta'_{j,k},\ \alpha_{i,j,k}^{(l),*} = \alpha'_{i,j,k}$
10. else $t = t + 1$
11. Repeat Steps 6-10 until the algorithm converges or $t = T^{\max}$
12. Set $l = l + 1$,
13. Repeat Steps 5-11 until $l = L$

14. $\{\delta_{i,j}^*, \beta_{j,k}^*, \alpha_{i,j,k}^*\} = \arg\min D^{(l)}(\delta_{i,j}^{(l),*}, \beta_{j,k}^{(l),*}, \alpha_{i,j,k}^{(l),*})$.

5 Simulation Results

In this section, we use MATLAB simulation software to evaluate and analyze the performance of the proposed algorithm. The simulation scenario consists of a single BS, multiple RUs and multiple SUs. The BS and users in the network are distributed in an area of 200 m × 200 m. The coordinates of the BS are (100 m, 100 m), and the positions of the users are randomly distributed. The number of users selected in the simulation is 8, the transmit power of the BS 26 dBm, the minimum transmission rate requirement of the RUs is set as 2 Mbit/s, and the power spectral density of the noise is set as −174 dBm/Hz, −160 dBm/Hz and −150 dBm/Hz.

Figure 2 shows the relationship between the total service delay and the number of iterations obtained from the algorithm proposed in this paper. In the simulation, the number of CHs is considered as 2 and 4, and the system transmission performance corresponding to different subchannel bandwidth is considered. It can be seen that the total service delay tends to converge within a small number of iterations, indicating the effectiveness of the proposed algorithm. Comparing the delay performance corresponding to different subchannel bandwidth, we can see the total service delay reduces as the subchannel bandwidth increases. In addition, comparing the service delay

performance corresponding to different number of CHs, it can be seen that when the number of CHs increases, the total service delay increases. This is because as the number of CHs increases, the traffic load of the transmission links between the BS and the CHs increases. As in general the link performance between the BS and the CHs may not be as good as the intra-cluster links, thus longer service delay might be resulted.

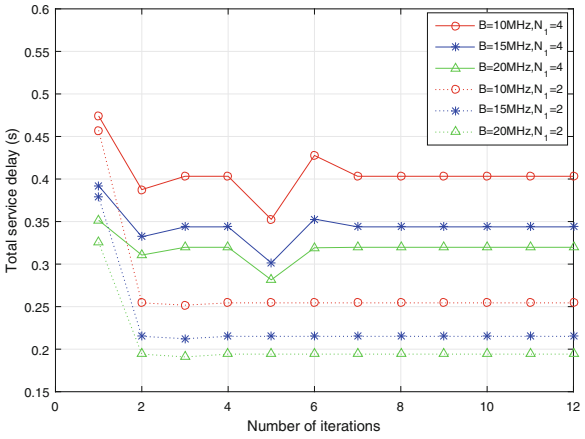

Fig. 2. Total service delay vs the number of iterations

Figure 3 shows the relationship between the total service delay of the users and the arrival rate of the BS obtained from the algorithm proposed in this paper. It can be seen that the total service delay of the users increases as the service arrival rate service increases. The reason is that when the service arrival rate service increases, larger number of packets are required to be transmitted, hence, longer queuing delay is resulted which causes longer total service delay in turn. Comparing the performance obtained from different average service rate, we can see that as the average service rate increases, the total service delay decreases which is benefited from shorter queuing delay. In addition, we can also observe that the increase in the number of CHs leads to an increase in the total service delay.

Figure 4 shows the relationship between total user service delay and traffic arrival rate under different minimum transmission rate limits. It can be seen that for relatively low minimum transmission rate requirement, lower total service delay of the users can be achieved, this is because to meet the minimum transmission rate requirement, a larger number of links might be qualified, thus offering higher flexibility in determining user association strategy and better service delay performance in turn.

Figure 5 shows total service delay versus subchannel bandwidth for different traffic arrival rates. For comparison, we also plot the performance of the algorithm proposed in [15]. It can be seen that given service arrival rate, the total service delay of the users decreases as the subchannel bandwidth increases. This is because the higher subchannel bandwidth offers a higher transmission rate and lower service delay in turn. We can also observe that the total service delay increases as the traffic arrival rate increases

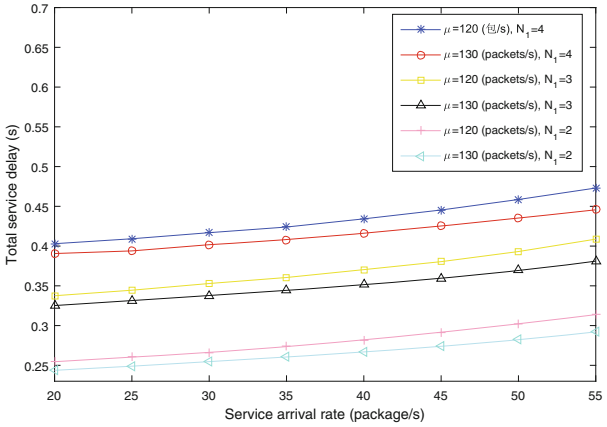

Fig. 3. Total service delay vs service arrival rate (different service rates)

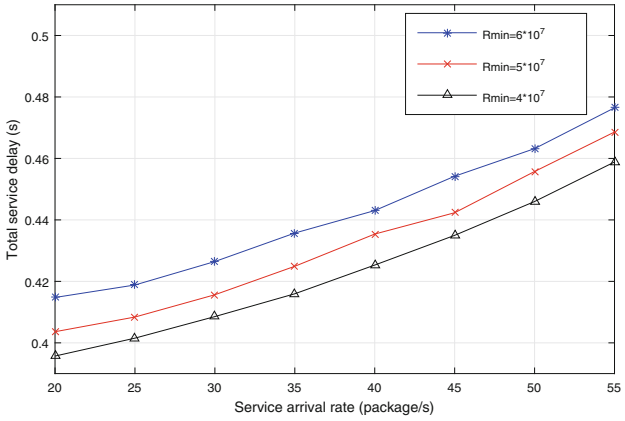

Fig. 4. Total service delay vs service arrival rate (different minimum transmission rate)

as larger amount of service results in longer queuing delay and longer service delay as well. Comparing the service delay performance obtained from our proposed algorithm and the algorithm proposed in [15], we can see that our proposed algorithm offers lower service delay than that proposed in [15]. The reason is that our proposed algorithm addresses joint optimization of user association and content placement and aims to achieve the optimal service delay, while the algorithm proposed in [15] is addressed in algorithm.

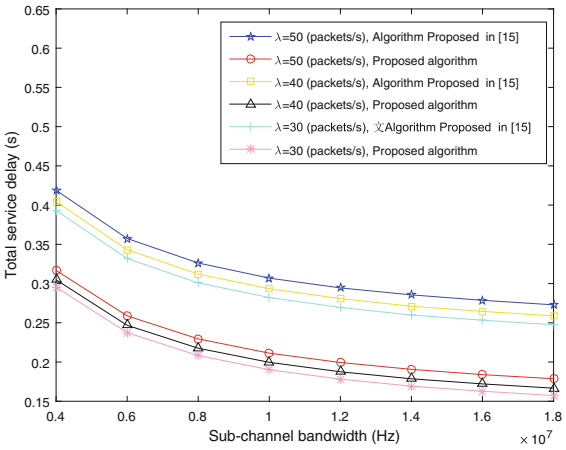

Fig. 5. Total service delay vs subchannel bandwidth

6 Conclusions

For the scenario of single-base cell cellular D2D communication system composed of multiple RUs and multiple SUs, this paper proposes a joint user association and content placement algorithm for cellular D2D communication based on delay optimization. In order to achieve the performance of most users, a clustering mechanism is proposed, which supports SU as the cluster head and supports the sharing of content by RU. Considering the constraints of cluster number, user association cluster head, cluster head cache capacity and transmission rate, a joint user association and content placement optimization model based on user's total service delay is established. In this paper, the Lagrange partial relaxation method is used to convert the original optimization problem into three sub-problems of convex optimization, and an iterative algorithm is proposed to jointly solve the sub-problems to obtain joint clustering optimization strategy and content placement optimization strategy. Finally, the proposed algorithm can realize the optimization of traffic transmission delay by MATLAB simulation.

References

1. Tehrani, M.N., Uysal, M., Yanikomeroglu, H.: Device-to-device communication in 5G cellular networks: challenges, solutions, and future directions. IEEE Commun. Mag. **52**(5), 86–92 (2014)
2. Asadi, A., Wang, Q., Mancuso, V.: A survey on device-to-device communication in cellular networks. IEEE Commun. Surv. Tutorials **16**(4), 1801–1819 (2014)
3. Fodor, G., Dahlman, E., Mildh, G., et al.: Design aspects of network assisted device-to-device communications. IEEE Commun. Mag. **50**(3), 170–177 (2012)
4. Zhu, H.: Radio resource allocation for OFDMA systems in high speed environments. IEEE J. Sel. Areas Commun. **30**(4), 748–759 (2012)

5. Klugel, M., Kellerer, W.: Leveraging the D2D-gain: resource efficiency based mode selection for device-to-device communication. In: IEEE Global Communications Conference (GLOBECOM), Washington, USA, pp. 1–7 (2017)
6. Wen, D., Yu, G., Xu, L.: Energy-efficient mode selection and power control for device-to-device communications. In: IEEE Wireless Communications and Networking Conference (WCNC), Doha, Qatar, pp. 1–7 (2016)
7. Penda, D.D., Liqun, F., Johansson, M.: Mode selection for energy efficient D2D communications in dynamic TDD systems. In: IEEE International Conference on Communications (ICC), London, UK, pp. 5404–5409 (2015)
8. Wang, K., Yu, F., Li, H.: Information-centric virtualized cellular networks with device-to-device communications. IEEE Trans. Veh. Technol. **65**(11), 9319–9329 (2016)
9. Pan, Y., Pan, C., Zhu, H., et al.: On consideration of content preference and sharing willingness in D2D assisted offloading. IEEE J. Sel. Areas Commun. **35**(4), 978–993 (2017)
10. Li, X., Ma, L., Shankaran, R., et al.: Joint mode selection and proportional fair scheduling for D2D communication. In: IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), Montreal, Canada, pp. 1–6 (2017)
11. Liberti, L., Pantelides, C.C.: An exact reformulation algorithm for large nonconvex NLPs involving bilinear terms. J. Glob. Optim. **36**(2), 161–189 (2006)
12. Boyd, S., Vandenberghe, L.: Convex optimization. Eur. J. Oper. Res. **170**(1), 326–327 (2006)
13. Huang, Y., Nasir, A.A., Durrani, S., et al.: Mode selection, resource allocation, and power control for D2D-enabled two-tier cellular network. IEEE Trans. Commun. **64**(8), 3534–3547 (2016)
14. Jiang, W., Feng, G., Qiu, S.: Optimal cooperative content caching and delivery policy for heterogeneous cellular networks. IEEE Trans. Mob. Comput. **16**(5), 1382–1393 (2017)
15. Ma, R., Xia, N., Chen, H.H., et al.: Mode selection, radio resource allocation, and power coordination in D2D communications. IEEE Wirel. Commun. **24**(3), 112–121 (2017)