UNIVERSITY OF WEST LONDON
The Career University

UWL REPOSITORY

repository.uwl.ac.uk

Caching deployment algorithm based on user preference in device-to-device networks

Fan, Hongmei, Zhang, Tiankui, Loo, Jonathan ORCID logoORCID: https://orcid.org/0000-0002-2197-8126 and Liu, Dantong (2018) Caching deployment algorithm based on user preference in device-to-device networks. In: IEEE Global Communications Conference (GLOBCOM) 2017, 04-08 Dec 2017, Singapore.

http://dx.doi.org/10.1109/GLOCOM.2017.8254692

This is the Accepted Version of the final output.

UWL repository link: https://repository.uwl.ac.uk/id/eprint/3640/

Alternative formats: If you require this document in an alternative format, please contact: open.research@uwl.ac.uk

Copyright:

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy: If you believe that this document breaches copyright, please contact us at open.research@uwl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Caching Deployment Algorithm Based On User Preference in Device-to-Device Networks

Hongmei Fan[1], Tiankui Zhang[1], Jonathan Loo[2], Dantong Liu[3]
[1]SICE, Beijing University of Posts and Telecommunications, Beijing, 100876, China
[2]School of Computing and Engineering, University of West London, United Kingdom
[3]Chief technology and Architecture office, Cisco Systems Inc. CA 95134, USA

***Abstract*—In cache enabled D2D communication networks, the cache space in a mobile terminal is relatively small compared with the huge amounts of multimedia contents. As such, a strategy for caching the diverse contents in a multiple cache-enabled mobile terminals, namely caching deployment, will have a substantial impact to network performance. In this paper, a user preference aware caching deployment algorithm is proposed for D2D caching networks. Firstly, based on the concept of the user preference, the definition of user interest similarity is given, in which it can be used to evaluate the similarity of user preferences. Then a content cache utility of a mobile terminal is defined by taking the communication coverage of this mobile terminal and the user interest similarity of its adjacent mobile terminals into consideration. The logarithmic utility maximization problem for caching deployment is formulated. Subsequently, we relax the logarithmic utility maximization problem, and obtain a low complexity near-optimal solution via dual decomposition method. The convergence of the proposed caching deployment algorithm is validated by simulation results. Compared with the existing caching placement methods, the proposed algorithm can achieve significant improvement on cache hit ratio, content access delay and traffic offloading gain.**

***Index Terms*—content caching, D2D, user preference, utility**

I. Introduction

Today's internet traffic is dominated by content distribution and retrieval. With the rapid explosion of the data volume and content diversity, it becomes challenging to deliver high quality service to the end user efficiently and securely. Content caching, a widely adopted content delivery technique in Internet for reducing network traffic load, has been exploited in fifth generation (5G) mobile networks. It has been proven that caching of popular content and pushing them close to consumers can significantly reduce the mobile traffic [1].

Device-to-device (D2D) communication has been regarded as another driving force behind the evolution into 5G. D2D caching networks, which take both advantages of caching and D2D communication technologies, have naturally set the stage for the 5G evolution [2]. In the D2D caching networks, the mobile terminals (MTs) equipped with storage space are used as caching nodes, and the MTs collaborative download and cache different parts of the same content from the serving base station (BS), and then share them by using D2D communications.

With a limited amount of storage on each MT, the main challenge is how cellular traffic can be maximally offloaded by using D2D communication to satisfy requests for content as well as to share messages between neighboring devices. A carefully designed caching deployment strategy would have a great impact on the network performance of the D2D caching networks. Some literatures have studied the caching deployment optimization problem [3]–[5]. In [3], a caching allocation scheme was proposed to enhance storage utilization for D2D networks, and the optimal storage assignment achieved trade-off between static caching and on-demand relaying. In [4], the authors studied the problem of maximizing cellular traffic offloading with D2D communication by selectively caching popular content locally, and exploring maximal matching for sender-receiver pairs. In [5], the authors optimized the content cache distribution considering the user's geographical location in the D2D network to improve the cache hitting probability.

These works utilize the D2D cache to achieve their goals by taking into account the channel state information, the popularity of the content, the available bandwidth resources, the data transmission rate, and the distribution of users. However, in the context of multimedia content distribution, especially in wireless social networks, the user's preference for content has a great impact on the cache system performance.

In [6], the authors pointed out that each data object would eventually be delivered to interested users. From users perspective, the closer of the storage location of the data object will result in less generated network traffic to access the data objects. With this in mind, the users interest preference will provide certain guidelines in selecting the caching location of the content replicas: it is beneficial to store the content replicas in the location much closer to the user who is interested in it. As such, the caching deployment strategy can be designed on the basis of the user preference.

In the context of D2D caching networks, [9] considered the difference of users preference and the selfishness nature of D2D users, and proposed a caching incentive scheme.

It is worth mentioning that user preference based caching strategies have been explored in content centric networking (CCN) recently. In a content centric multi-hop wireless network, a caching placement strategy based on user interest was proposed [10]. A cooperative caching strategy was proposed in CCN [11], which took user preference, node importance, and cache replacement rate into account when making the caching decision. Nevertheless, the works in CCN pay more attention to the online and on-path caching decision design, which

cannot obtain the overall network performance optimization.

Although the works in [9]–[11] laid a good foundation in integrating user's interest preference to the caching strategy design, the effect of the similarity of the users interest preferences on the caching strategy design is less well understood. In the caching strategy design, if the MT caches some specific contents which may be interested by the adjacent MTs, the cache space utilization can be improved. Hence, the content caching of a MT should not only consider the user preference itself, but also take account of the user interest similarity on the contents of adjacent MTs. Our work fills the gap by carefully considering the user interest similarity and the D2D transmission coverage region when defining the content cache utility, thereby improving the network performance via the caching deployment problem optimization.

In this paper, we propose a user preference aware caching deployment algorithm for the D2D caching networks. The content cache utility of each MT is defined to measure the caching utilization, which takes both the user preference and the transmission coverage region into consideration. The logarithmic cache utility maximization problem is introduced for the caching performance optimization. Then a near-optimal solution is obtained by dual decomposition method. This solution provides a feasible, efficient and low-overhead algorithm for implementation in D2D caching networks. The simulation results show that the proposed algorithm can converge to the maximization solution in a few iterations and achieve significant performance on cache hit ratio, content access delay and traffic offloading gain.

The rest of the paper is organized as follows: Section II presents the system model. Section III defines user interest similarity, and the content cache utility. In Section IV, we propose the content cache utility optimization problem and the near-optimal algorithm. Section V evaluates the performance of the proposed algorithm. Lastly, the conclusion is highlighted in Section VI.

II. System Model

A. Network model

The system model of this paper is illustrated in Fig. 1. A single macrocell is considered, where a macro BS serves N uniformly distributed D2D users. D2D communication utilizes orthogonal frequency, so there is no co-channel interference between D2D communications links. The bandwidth of each orthogonal frequency channel is B. Each MT has a cache space which is able to cache up to S contents. The popular contents are assumed to have the same data size, and the data volume of each content is v.

With each user node as the center, we calculate the number of neighbor nodes around each nodes communication coverage, and let Φ_n denotes the set of neighbors of user n in its communication coverage. The user n can communicate and share the content directly with his neighbor MTs through D2D communication link if user n cached a content. For a given content, which MT to save the content replica in an overlap region of multiple MTs will be decide by the caching replacement strategy.

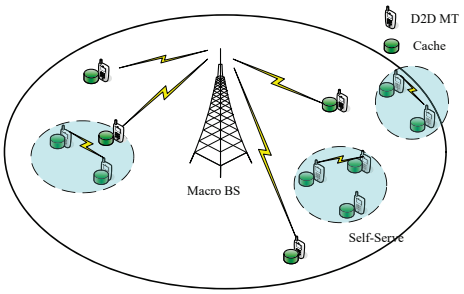

Figure 1. D2D caching networks.

When user n is communicated with user n', the data rate from user n' to user n is

$$c_{nn'} = B \log \left(1 + \frac{g_{n'n} p_n'^{tx}}{\sigma^2}\right), \quad (1)$$

when user n is communicated with BS, the data rate from its severing BS to user n is

$$c_{n\_BS} = B \log \left(1 + \frac{g_{BS\_n} p_{BS}^{tx}}{\sigma^2}\right), \quad (2)$$

where σ^2 is additive white Gaussian noise power, $p_n'^{tx}$ is the maximum transmit power of user n', p_{BS}^{tx} is the maximum transmit power of BS, and $g_{n'n}$ is the channel gain between n' and n, and g_{BS_n} is the channel gain between BS and n, which including the path-loss, shadow fading and multipath fading.

B. Caching model

Taking into account the diversity of contents within the D2D network, this paper assumes that all M contents which having the highest popularity are cached by at least one MT according to certain criteria.

In our caching model, a user can be a content requester and a content provider. If there exists a complete or partial copy of content m in its own cache, the request is fulfilled with no delay and without the need to establish a communication link. Otherwise, the user broadcasts a request message for the content m to the neighbor MTs within its coverage, if the user can find the requested file from a MT's cache space within its D2D transmission range, then it can establish a D2D communication link and obtain the content. If the user cannot find the requested content neither in its own cache nor its proximity users, it needs to download the file from the BS.

III. Cache Utility Function

In this paper, the user preference φ_{mn} is defined as the interest of user n to the content m based on cousin theory [12]. Further, the interest similarity function is defined to characterize the interest similarity among users. We use a simple model to capture the user interest similarity of real social networks [13]. Since φ_m is within the segment $[0, 1]$, the interest similarity between user n and user n' is defined as the Euclidean distance on the wrapped segment,

$$\varphi_m(n, n') = \min\left\{\left|\varphi_{mn} - \varphi_{mn'}\right|, 1 - \left|\varphi_{mn} - \varphi_{mn'}\right|\right\}. \quad (3)$$

The small distance between φ_{mn} and $\varphi_{mn'}$ is, the larger interest similarity of user n and n' on the content m is. A

larger interest similarity between two users indicates that the more likely content cached in one user is requested by the another user.

Then, we define the cache utility function of a user considering both the user preference and the D2D transmission coverage region. In the D2D communication coverage region of a MT, the more neighbor users, the higher the possibility of content sharing of the cached content, and the larger the caching utility of this MT.

As described above, $\varphi_m(n,n')$ represents the interest similarity between user n and user n' for the conetent m. Besides that, we let $d(n,n')$ represents the physical distance between the two users, and let Φ_n denotes the set of neighbors of user n in its communication range. Therefore, the cache utility per unit cache space of user n caching content m is defined as,

$$u_{mn} = \sum_{n' \in \Phi_n} \left[\varphi_m(n,n')^{-\alpha} \cdot d(n,n')^{-\beta} \right], \tag{4}$$

where α and β are the weighting factors of the user interest similarity and the user physical distance.

IV. Caching Deployment Algorithm

The goal of this paper is to optimize the cache utility of the whole network to obtain the caching deployment algorithm.

A. Problem Formulation

We define a caching index $x_{mn} = 1$ of user n and content m, indicating that a portion of (or entire) content m is cached in user n, otherwise $x_{mn} = 0$.

Assumption 1: user n can cache a portion of content m. This assumption is practical and necessary, because a user may have multiple interested content to be cached and the cache space in a MT is relative small compared with the data volume of the multiple contents.

Suppose the cache space of each MT is S, and the cache space allocated by the MT n for caching parts of the content m is $S\left(\sum_{m=1}^{M} x_{mn}\right)^{-1}$, namely equal cache space allocation.

Logarithmic utility function in particular is a very common choice of utility function [14], which naturally achieves some level of utility fairness among the contents. So we use a logarithmic utility function in the cache utility maximization problem. The optimization problem can be written as,

$$\begin{aligned} \mathbf{P1}: \max_{x} & \sum_{m=1}^{M} \left(\sum_{n=1}^{N} x_{mn} \log \left(\frac{Su_{mn}}{\sum_{m=1}^{M} x_{mn}} \right) \right) \\ \text{s.t. } & \text{C1}: \sum_{n=1}^{N} x_{mn} = 1,\ \forall m \in \{1,\cdots,M\}, \\ & \text{C2}: x_{mn} \in [0,1],\ \forall m \in \{1,\cdots,M\}, \\ & \quad \text{and } \forall n \in \{1,\cdots,N\}. \end{aligned} \tag{5}$$

The problem $\mathbf{P1}$ is combinatorial due to the binary variable x_{mn}, the complexity of the brute force search is $\Theta\left((N)^M\right)$, where N and M denote the number of MTs and number of contents, respectively. The computation is essentially impossible for even a modest-sized cellular network.

To overcome this, we give the **Assumption 2** as following to allow one content to be cached in multiple MTs.

Assumption 2: one content can be cached in multiple MTs simultaneously. This assumption may require more overhead to implement, but it is a practical method in D2D caching networks, since the multiple MTs can collaborative download and cache some large volume contents.

In the following, we provide a physical relaxation of C2 in (5) as $0 \le x_{mn} \le 1$. With this physical relaxation, the indicators x_{mn} can take on any real value in $[0,1]$, representing that one content can be cached portion in more than one MT, which follows the **Assumption 1** as well. So optimization problem can be expressed as,

$$\begin{aligned} \mathbf{P2}: \max_{x} & \sum_{m=1}^{M} \left(\sum_{n=1}^{N} x_{mn} \log \left(\frac{Su_{mn}}{\sum_{m=1}^{M} x_{mn}} \right) \right) \\ \text{s.t. } & \text{C1}, \\ & \text{C3}: 0 \le x_{mn} \le 1,\ \forall m \in \{1,\cdots,M\}, \\ & \quad \text{and } \forall n \in \{1,\cdots,N\}. \end{aligned} \tag{6}$$

To solve the convex optimization (6), the global network information is necessary, which requires a centralized controller for caching deployment and coordination between MTs.

B. Dual decomposition

From (6), we have $\log\left(Su_{mn}\left(\sum_{m=1}^{M} x_{mn}\right)^{-1}\right) = \log(Su_{mn}) - \log\left(\sum_{m=1}^{M} x_{mn}\right)$. We introduce a new variable $M_n = \sum_{m=1}^{M} x_{mn}$ representing the number of contents cached in the MT n. Then the problem of (6) can be rewritten as,

$$\begin{aligned} \mathbf{P2}': \max_{x} & \sum_{m=1}^{M}\sum_{n=1}^{N} x_{mn} \log(Su_{mn}) - \sum_{m=1}^{M}\sum_{n=1}^{N} x_{mn} \log M_n \\ \text{s.t. } & \text{C1, C3}, \\ & \text{C4}: \ M_n = \sum_{m=1}^{M} x_{mn}, \\ & \text{C5}: \ M_n \le M. \end{aligned} \tag{7}$$

The redundant constraint $M_n \le M$ is added for the analysis of convergence of the proposed algorithm which represent that the number of contents cached in user n is less than the maximum number of contents cached in user n. The only coupling constraint is $M_n = \sum_{m=1}^{M} x_{mn}$ in problem (7). This motivates us to turn to the Lagrangian dual decomposition method whereby a Lagrange multiplier λ is introduced to relax the coupled constraint. The dual problem is,

$$D: \min_{\lambda} D(\lambda) = f_x(\lambda) + g_{M_n}(\lambda), \tag{8}$$

in which,

$$\begin{aligned} f_x(\lambda) = \max_{x} & \sum_{m=1}^{M}\sum_{n=1}^{N} x_{mn} \log(Su_{mn} - \lambda_n) \\ \text{s.t. } & \text{C1, C3}, \end{aligned} \tag{9}$$

and

$$g_{M_n}(\lambda) = \max_{M_n \le M} \sum_{n=1}^{M} M_n (\lambda_n - \log M_n). \quad (10)$$

The primal problem (7) can be equivalently solved by the dual problem (8). Denoting $x_{mn}(\lambda^*)$ as the maximizer of the first sub-problem (9) and $M_n(\lambda^*)$ as the maximizer of the second sub-problem (10). There exists a dual optimal λ^* such that $x_{mn}(\lambda^*)$ and $M_n(\lambda^*)$ are the primal optimal. Therefore, given the dual optimal λ^* , we can get the primal optimal solution by solving the decoupled inner maximization problems of (9) and (10) separately.

C. Algorithm procedure

The procedure of the dual problem solution is executed by the macro BS. We assume that the all information of caching utility x_{mn} is known by the macro BS.

The outer problem is solved by the gradient projection method [15], where the Lagrange multiplier λ is updated in the opposite direction to the gradient $\nabla D(\lambda)$

$$\frac{\partial D(\lambda)}{\partial(\lambda_n)} = M_n - \sum_{m=1}^{M} x_{mn}. \quad (11)$$

Evaluating the gradient of the dual objective function (8) requires us to solve the inner maximization problem, which has been decomposed into two sub-problems f and g. These sub-problems are solved by **Algorithm 1** as follows.

Algorithm 1 Caching deployment algorithm

Set t as the iteration index, and define a caching index matrix $X = \{x_{mn}\}_{M\times N}$. A small positive number ε is predefined as the convergence constant.

Initialization: $t = 0$, $x_{mn} = 0$, and the macro BS generates a random multiplier λ_n for each MT.

Iteration: in the t^{th} iteration of gradient projection algorithm for the content m, the procedure is as following,

Step 1: the macro BS obtain the MT n satisfies $n^* = \arg\max_n (\log(Su_{mn}) - \lambda_n(t))$; then set $x_{mn^*} > 0$ and update $M_n^*(t+1) = \sum_{m=1}^{M} x_{mn^*}$;

Step 2: the macro BS updates the values of $M_n(t+1)$ according to the problem (10), we set its gradient to be 0 with the constraint $M_n \le M$, i.e., $\lambda_n - 1 - \log M_n = 0$, then we have, $M_n = e^{(\lambda_n(t)-1)}$, then the value of M_n is updated by

$$M_n(t+1) = \min\left\{M, e^{(\lambda_n(t)-1)}\right\}. \quad (12)$$

Step 3: the macro BS updates the Lagrange multiplier value $\lambda_n(t+1)$ by the following method,

$$\lambda_n(t+1) = \lambda_n(t) - \delta(t)\left(M_n(t) - \sum_m x_{mn}(t)\right), \quad (13)$$

where $\delta(t) > 0$ is a dynamically chosen step size sequence based on some suitable estimates.

Convergence Judgment: when $|D(t+1) - D(t)| \le \varepsilon$, the iteration will stop and the caching deployment result is given by caching index matrix X; otherwise, it will go to the next iteration.

It can be proved that the **Algorithm 1** is guaranteed to converge to a near-optimal solution [15]. At each iteration, the complexity of the proposed algorithm is $O(MN)$. The gradient method converges fast generally, and thus the number of iterations is a small number (less than 10 in the simulation). Meanwhile, we assume that the convergence of the proposed algorithm is faster enough compared with the timescale of the cached contents updated and replacement.

V. Performance Evaluation

In the simulation, a macro BS is deployed at the center of the cell and N users are uniformly randomly distributed in the cell, and can communicate with any neighbor users in each user own coverage. We assume the storage capacity of each user are equal and the initial states are empty. The popularity of M contents follows a *Zipf*-like distribution as previous studies [10], and the content size v is set to 1024 bytes. We also assume that the macro BS is aware of all the users preferences, i.e. $\{\varphi_{mn}\}$ is a common knowledge within the network.

Here we assume that the macro BS has all the user requested contents by downloading from the Internet, the bandwidth for D2D communication is 20 MHz, and D2D communication links use orthogonal frequency resources. The detailed simulation parameters are given in Table I.

TABLE I: SIMULATION PARAMETERS

Parameters	Value
Bandwidth	20MHz
Cell radius	500m
Maximum transmission range of D2D links	50m
Maximum transmit power of BS	46dBm
Transmit power of D2D	30dBm
Log-normal shadowing fading	10dB
Noise power	-174dBm/Hz
Number of users	10∼100
Number of contents	10∼100

We compare the performance of the following caching schemes:

Preference Aware Caching (PAC): The proposed caching placement algorithm via dual decomposition.In the simulation, the parameters related with PAC are setting as, α= 1, β= 1, ε= 0.1, δ= 2, θ= 0.8, ρ= 1.6 and μ=M.

Random Complete Caching (RCC): In this caching scheme MTs cache data items randomly and then other MTs choose to get the replica from the nearest cache MT or from the data source. In this caching scheme, only the cache MTs choose their strategies randomly.

The performance criteria considered are the average content access delay, cache hit ratio, offloading ratio, and the content cache utility. The detailed definitions are given as following.

Average content access delay: The average content access delay is the service delay of data request within the simulation period, which is calculated as,

$$\tau = \frac{1}{MN}\sum_{n=1}^{N}\sum_{m=1}^{M}\left(\frac{v\varphi_{mn}}{c_{nn'}x_{mn} + c_{n\_BS}(1 - x_{mn})}\right), \quad (14)$$

where, $c_{nn'}$ and c_{n_BS} are the data rate from D2D user and BS calculated according to (1) and (2).

Cache hit ratio: The percentage of queries of data item being satisfied within its own communication range or its own cache space during the simulation period. The cache hit ratio is calculated as,

$$\eta = \frac{\sum_{n}^{N}\sum_{m}^{M}\left(1-e^{-d^2}\right)\varphi_{mn}x_{mn}}{\sum_{n}^{N}\sum_{m}^{M}\varphi_{mn}}. \quad (15)$$

Offloading ratio: This is the ratio of the amount of data offloaded by the D2D content delivery to the total amount of the request data in the cell, which reflects the ability of offloading the traffic of BS. Considering the actual application scenario, the data volume of all the contents is much larger than the cache space available to each MT, so the expression of the offloading ratio is,

$$h = \frac{\sum_{n}^{N}\sum_{m}^{M}\left(\left(1-e^{-d^2}\right)\varphi_{mn}x_{mn}S\left(\sum_{m=1}^{M}x_{mn}\right)^{-1}\right)}{v\sum_{n}^{N}\sum_{m}^{M}\varphi_{mn}}. \quad (16)$$

Firstly, the convergence of the proposed PAC algorithm is verified in the Fig. 2. In the simulation, the cache space size of each MT is set as S = 8, and the number of contents is M = 10. It can be observed that the PAC can converge to the approximate optimal solution within 10 times iterations both in the cases of 20 and 40 MTs. When N = 40, the fluctuations before convergence are relatively larger compared with the case of N=20. This is because when there are the more users, there will be more the state of the caching, consequently the proposed algorithm needs to take more time to choose the best fitted MT to cache the contents before reaching a steady state.

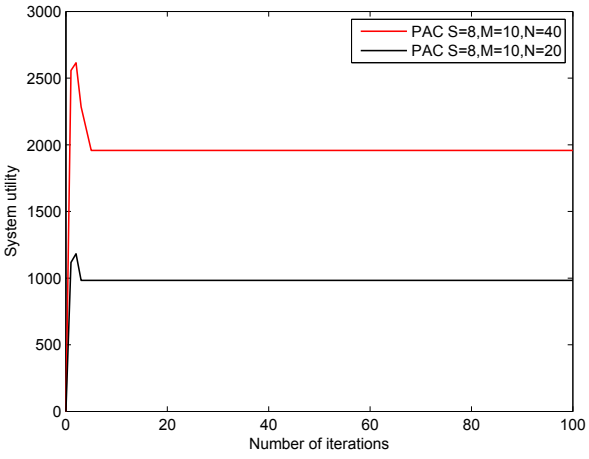

Figure 2. Convergence validation of the PAC.

Then we verify the performance of the caching schemes with varying number of contents from Fig. 3-Fig. 6. Here we set the parameter to: the number of users is N=100, the number of contents changes from 20 to 100, and the number of contents S = 4 and S = 8, respectively. We can see that all the performance of the proposed PAC is significantly improved compared to the RCC.

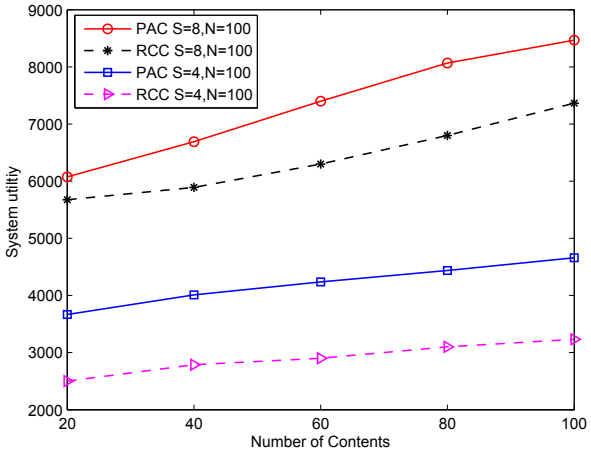

Figure 3. System utility with varying number of contents.

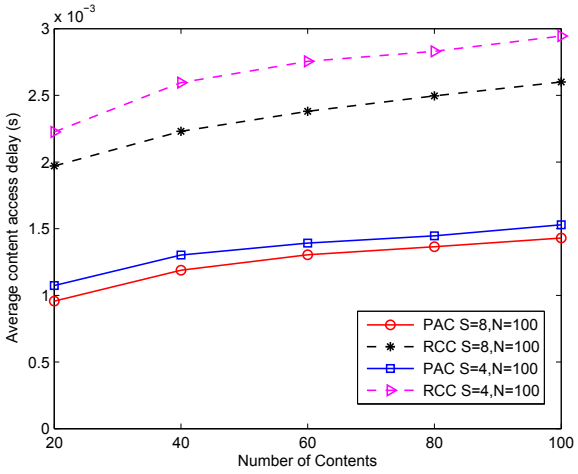

Figure 4. Average content access delay with varying number of contents.

Fig. 3 shows that the total system utility increases as the number of users increasing, and the utility value of the proposed PAC is always higher than that of the RCC. This is due to the fact that with the increase of the number of contents, the distribution of popular content is more and more dispersed, and the user preferences for contents are also increasingly scattered. RCC randomly select contents without considering the user preference, whereas the proposed PAC considers the user preference to maximize the caching utility of the users.

The comparison of the average content access delay with varying number of contents is given in Fig. 4. We can see that the proposed PAC achieves much lower content access delay compared to the RCC. Meanwhile, the average content access delay increases with the increasing of the number of contents. This is due to the fact that with the increase of the number of contents, the D2D MTs with limited cache space cannot cache

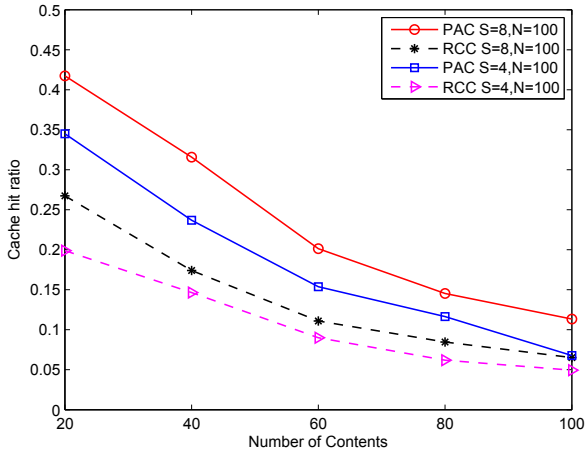

Figure 5. Cache hit ratio with varying number of contents.

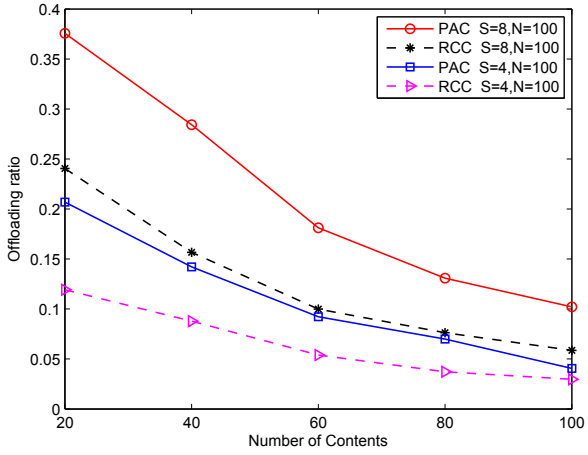

Figure 6. offloading ratio with varying number of contents.

all the contents which requested by the neighbor users, which leads that macro BS will provide more contents to the users.

Fig. 5 compares the cache hit ratio of the PAC and RCC for different numbers of contents. The cache hit ratio of the proposed PAC decreases as the number of contents increasing. The reason is that with the increasing of the number of contents, the users cannot cache all contents with limited cache space. The cache hit ratio performance gain of the PAC reaches up to 61% and 75% relative to the RCC when S=8 and S=4, respectively. The gain increases as the storage capacity increases, since higher capacity allows the MTs to cache more popular contents.

From Fig. 6, we observed that the offloading ratio of the proposed PAC has great improvement compared that of the RCC, especially when the case space size is large. The results also show that, the offloading ratio decreases with the increasing of the content number, because more contents requests cannot be responded.

VI. Conclusion

In this paper, we have proposed a user preference aware caching deployment algorithm. The proposed algorithm measures the content caching utilization taking account of both the user preference and the transmission coverage region. By doing so, the proposed algorithm would be able to cache specific contents that match the user preferences that may also be interested by the adjacent nodes at unpopulated region. Beyond that, we have introduced a cache utility function with the aim to maximize cache utility in order to enhance the possibility of content sharing of among the multiple MTs. The proposed algorithm has obtained the near-optimal performance of the caching deployment, which can be used as the benchmark for the online caching strategy design.

References

[1] J. Erman, A. Gerber, M. Hajiaghayi, D. Pei, S. Sen, and O. Spatscheck, "To cache or not to cache: The 3G case," *IEEE Internet Computing*, vol. 15, no. 2, pp. 27–34, March 2011.

[2] N. Golrezaei, A. F. Molisch, A. G. Dimakis, and G. Caire, "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution," *IEEE Communications Magazine*, vol. 51, no. 4, pp. 142–149, April 2013.

[3] W. Wang, X. Wu, L. Xie, and S. Lu, "Joint storage assignment for D2D offloading systems," *Computer Communications*, vol. 83, pp. 45–55, 2016.

[4] J. Jiang, S. Zhang, B. Li, and B. Li, "Maximized cellular traffic offloading via device-to-device content sharing," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 1, pp. 82–91, Jan 2016.

[5] D. Malak, M. Al-Shalash, and J. G. Andrews, "Optimizing the spatial content caching distribution for device-to-device communications," in *2016 IEEE International Symposium on Information Theory (ISIT)*, July 2016, pp. 280–284.

[6] K. C. Chen, M. Chiang, and H. V. Poor, "From technological networks to social networks," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 9, pp. 548–572, September 2013.

[7] J. Lu, D. Wu, M. Mao, W. Wang, and G. Zhang, "Recommender system application developments: a survey," *Decision Support Systems*, vol. 74, pp. 12–32, 2015.

[8] Y. Guo, L. Duan, and R. Zhang, "Cooperative local caching and file sharing under heterogeneous file preferences," in *2016 IEEE International Conference on Communications (ICC)*, May 2016, pp. 1–6.

[9] Z. Chen, Y. Liu, B. Zhou, and M. Tao, "Caching incentive design in wireless D2D networks: A stackelberg game approach," in *2016 IEEE International Conference on Communications (ICC)*, May 2016, pp. 1–6.

[10] J. Iqbal and P. Giaccone, "Interest-based cooperative caching in multi-hop wireless networks," in *2013 IEEE Globecom Workshops (GC Wkshps)*, Dec 2013, pp. 617–622.

[11] L. Wu, T. Zhang, X. Xu, Z. Zeng, and Y. Liu, "Grey relational analysis based cross-layer caching for content centric networking," in *2015 IEEE/CIC International Conference on Communications in China (ICCC)*, Nov 2015, pp. 1–6.

[12] X. Han, L. Wang, S. Park, Á. Cuevas, and N. Crespi, "Alike people, alike interests? a large-scale study on interest similarity in social networks," in *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*, Aug 2014, pp. 491–496.

[13] M. Boguñá, R. Pastor-Satorras, A. Díaz-Guilera, and A. Arenas, "Models of social networks based on social distance attachment," *Physical review E*, vol. 70, no. 5, p. 056122, 2004.

[14] S. Stanczak, M. Wiczanowski, and H. Boche, *Fundamentals of resource allocation in wireless networks: theory and algorithms*. Springer Science & Business Media, 2009, vol. 3.

[15] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 6, pp. 2706–2716, June 2013.