Cooperative Caching and Delivery Algorithm Based on Content Access Patterns at Network Edge

Lintao Yang[1], Yanqiu Chen[2], Luqi Li[2], and Hao Jiang[2](✉)

[1] College of Physical Science and Technology, Central China Normal University, Wuhan 430079, China

[2] School of Electronic Information, Wuhan University, Wuhan 430072, China
jh@whu.edu.cn

Abstract. Mobile network performance and user Quality of Experience (QoE) will be negatively affected by the explosion of mobile data traffic. Recent research has focused on local caching at the wireless edge, as motivated by the 80/20 rule regarding content popularity. By caching popular contents at base stations (BSs), backhaul congestion and content access latency can be dramatically reduced. To address the limited storage size of BSs, an algorithm optimizing cooperative caching has been highlighted. Contents requested by mobile users that cannot be obtained locally could be transferred by cooperative BSs. In this paper, we propose a cooperative caching algorithm based on BS content access patterns. We use tensor decompositions with distance constraint to analyze interaction between users, contents and base stations. Thus, BSs with small geographical distances and similar content access patterns constitute a cooperative caching domain. Simulation results based on a real dataset of usage detail records (UDRs) demonstrate the superior performance and promising practical gains in caching of the proposed caching method compared to user clustering and BS clustering.

Keywords: Cooperative caching · Mobile Internet · Multi-aspect data and analysis · Network edge · Tensor decomposition

1 Introduction

In recent years, rapid ubiquity of advanced mobile applications with huge bandwidth requirements has dramatically increased mobile and Wi-Fi traffic. The Cisco Visual Networking Index (VNI) predicts that mobile data traffic will increase at a compound annual growth rate of 47% resulting in an increase to 49 exabytes monthly by 2021, which is a 7-fold increase over mobile data traffic in 2016. Globally, the proportion of smart devices and connections will increase to 75% in 2021, compared to 46% in 2016. As the explosive growth of network traffic, traditional core network architecture cannot accommodate rapidly increasing

V. C. M. Leung et al. (Eds.): 5GWN 2019, LNICST 278, pp. 99–123, 2019.
https://doi.org/10.1007/978-3-030-17513-9_8

user demand generated by the explosive growth of network traffic, due to long latency and the heavy burden on backhaul links.

Mobile edge caching [1,2] has been identified as one of the most disruptive enablers for 5G networks by the 5G Infrastructure Public Private Partnership, both to reduce content access latency and to alleviate backhaul congestion by relocating computing and storage units closer to the edge of the network [3]. This approach better accommodates proximal user demand while eliminating redundant transmissions from the remote sources [4].

To meet the ever-increasing demand for resource utilization and Quality of Experience (QoE), efficient cooperative content caching and delivery algorithm is suggested for cooperative caching between users and BSs. In most previous research on cooperative caching, cooperative base stations (BSs) are grouped based on placement, without considering content access behavior at the network edge or the interest distribution of mobile users. Thus, the cache hit ratio is not maximized.

With the popularity of the mobile Internet, the diversity of user content access behavior reflects to the heterogeneity of users [5]. With certain numbers of BSs, the distribution of mobile user interest and the effect of spatial-temporal information are predictable [6]. Therefore, we can apply cooperative caching based on user access patterns at the network edge to facilitate familiar content access behavior in clusters of BSs and achieve more efficient utilization of network resources. In next-generation cellular networks, the distance between small cells decreases and while physical layer technology is already at the boundary of Shannon capacity [7,8]. The proposed method would decrease the cost of collaboration, making cooperation caching based on user access patterns reasonable and feasible.

This paper proposes a cooperative caching algorithm based on user access patterns at the network edge. Local caching by cooperative base stations at the network edge constitutes a cooperative domain, in which base stations share contents. The cooperative domain is determined by clustering based on multiple aspects of mobile user content requests at the network edge. We have designed an efficient content placement and delivery algorithm that maximizes the cache hit ratios. Based on a dataset of real usage detail records (UDRs), results demonstrate that the proposed algorithm achieves a higher hit ratio while controlling the cooperative cost and improving QoE.

2 Related Work

Cooperative caching eases the traffic in the core network and reduces the overall download time, hence enhancing user perceptions of QoE [9]. Some studies have focused on cooperation between caching nodes, grouping them to improve the efficiency of network resource utilization and QoE.

Chen et al. [10,11] proposed a cluster-centric small cell network with a combined cooperative caching and delivery algorithm. Small base stations (SBSs) were grouped into disjoint clusters based on geographic position information, in which in-cluster cache space was utilized as an entity. Wang et al. [12] proposed a beamforming scheme that coordinates multiple remote radio heads (RRHs) in C-RAN to improve the quality of experience (QoE) of users by maximizing their aggregate weighted quality of service (QoS). Hu et al. [13] presented a general framework to model the video diffusion among mobile users and user QoS of the MSVS service over the wireless infrastructure. Li et al. [14] focused on the cooperative cell caching for future mobile networks, where each cell (e.g., base station) can cache popular contents for improving QoS. Fan et al. [15] proposed a clustering-based downlink resource allocation algorithm to allocate downlink spectrum resources in small cell networks. Yan et al. [16] proposed a hierarchical clustering-based caching strategy that improved caching efficiency by using cooperative caching for BS communication.

Another well-known strategy is cooperative caching based on user clustering. In [17,18], the users within the network were clustered according to their content popularity distribution and caching was executed accordingly to maximize the hit ratios. Unlike clustering algorithms based on user interest distribution, in [19], users were grouped based on their locations. To maximize the cache hit ratios, BSs used joint caching and delivery policies for users within communication range.

Cooperative content placement and delivery algorithms for known cooperative BSs is a popular research topic in cooperative caching. These studies generally assume or emulate the cooperation relationships of BSs directly, using this as basis of efficient algorithms designed to improve QoE. In [20], Scalable Video Coding (SVC) with cooperative caching was used to enable caching and serve sliced video layers that can serve different bitrates to improve utilization of caching resources. In [21,22], the authors developed light-weight cooperative cache management algorithms based on a heterogeneous cellular network (HetNet). These algorithms are considered promising architectural techniques for 5G as they maximize the traffic volume served by caching while minimizing the bandwidth cost.

The prior studies explore cooperation caching design content replacement and delivery algorithms for cooperative BSs to maximize QoE. Though diverse results have been acquired, the determination of cooperative relationship must be revised. When the cooperative relationship of base stations is directly assumed, cooperative BSs are not grouped and only a few base stations are considered in the simulation. Conversely, in cooperation caching, in which BSs are grouped based on geographical positions, user mobile access behavior is ignored. Cooperation caching, that is, grouping users based on interest distribution or placement, should inform the cooperative relationship with BSs. However, this approach generates suboptimal results in terms of precision.

Inspired by the preceding studies, we assert that grouping BSs into clusters is more direct because the cooperation caching occurs on the caching unit. Cooperative BSs with close geographical distances access similar content, thus efficiently utilizing caching resources. The method proposed in this paper can cluster BSs based on mobile user content access behavior at the network edge by considering multiple aspects to determine a suitable set of cooperative BSs.

3 User Behaviour Analysis

Before introducing the cooperative caching algorithm, we will analyze user access patterns and show the feasibility of our method.

We considered a real dataset of UDRs obtained from a mobile network operator in Jinhua, China. The dataset contained the data access records of 1.6 million mobile phone users over 23 days, covering 8,845 base stations, and involving up to 172,324 contents.

The dataset, which was stored in MySQL, contained information about the content access behavior of mobile users and as described by the fields listed in Table 1. Note that user information privacy was preserved by encrypting sensitive data.

Table 1. Data of the usage detail records.

Fields	Description
UID	An encrypted telephone number indicating a mobile user
Time_start	The time that a user begins the content request
Time_end	The time that a user ends the content request
LAC & CID	The base station providing content resources
URL	The identity of the content access

The UDR dataset also contained the geographical position information provided by the corresponding relationship of local area code (LAC) and cell ID (CID) and position as depicted in Table 2.

Table 2. Description of BSs geographical position information.

Fields	Description
LAC & CID	The base station providing content resources
Longitude	Longitude of BSs
Latitude	Latitude of BSs

First, we consider content popularity. Various types of content are available to mobile users. But, owing to the individuality reflected by user preferences and the demonstration effect, the popularity of diverse content differs among users.

Figure 1 shows the diversity content popularity distribution. Content popularity reflects average interests of multiple users.

From Fig. 1(b), we observe that within 23 days, the content with a total visit frequency greater than 100 accounted for only 4.659% of traffic, and the content with a total visit frequency greater than 1000 accounted for only 1.166% of traffic. This demonstrates that content popularity follows the power-law distribution. Fewer popular contents are accessed by larger numbers of users. This imbalance in content popularity follows the Pareto principle, that is, roughly 80% of traffic is attributed to 20% of the content. Thus, it is sensible to cache content with high popularity at base stations, as this practice would consume little cache space relative to the vast majority of user interests covered, thereby greatly improving the utilization of cache resources and enhancing QoE.

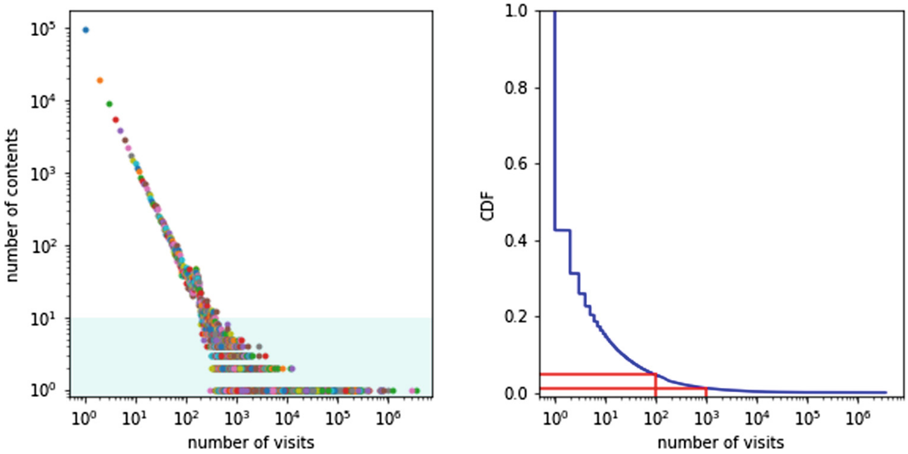

Fig. 1. Content popularity: (a) statistical distribution and (b) cumulative distribution function (CDF).

From Fig. 1 also note that the content popularity distributions exhibited a long-tail effect, and the second gradient of access, in which access frequency was less than 100, contained many types of content. Although the number of visits did not reach the maximum, the total number of visits and the coverage of the mobile user group cannot be ignored [23]. The long tail effect is fully utilized in recommender systems [24]. Based on this observation, we suggest that when considering edge caching, we should increase content diversity as much as possible while still guaranteeing the popularity of cached content. Thus, we maximize the limited cache size to satisfy more users' requests locally. The second consideration concerns user behavior. Mobile networks offer a broad range of content, available anywhere at any time, without any cost difference. But, whether every user will access the majority of network content is questionable. We use the UDR dataset to answer this question. Per Fig. 2, a total of 84.835% of mobile users had fewer than 10 visits during the 23-day observation period. This indicates that most mobile users were interested in a low variety of content. From Fig. 2, we learn that every user prefers certain content, but that less popular content is accessed by most mobile users. Hence, there must be an overlap of the individual user preferences, making the sharing of cached content feasible.

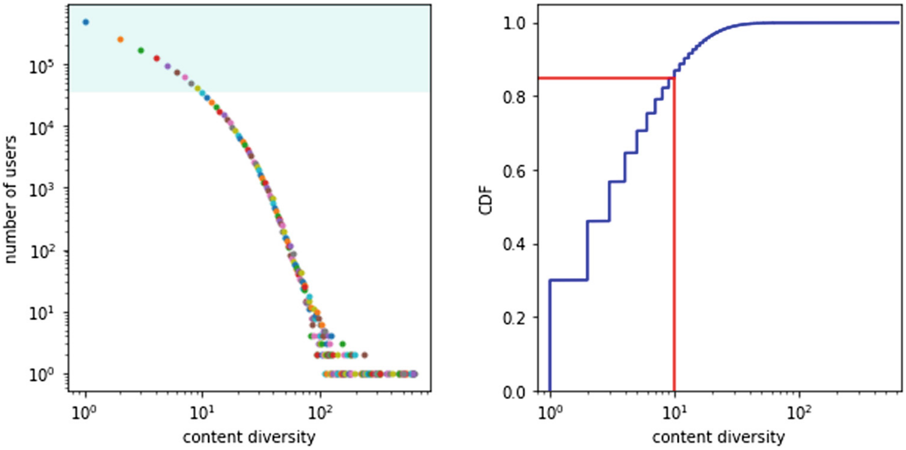

Fig. 2. Mobile user preference: (a) statistical distribution and (b) cumulative distribution function (CDF).

Regarding the content to be cached at BSs, we explore whether there are differences or similarities in the distribution of content requested by mobile users under different base stations. For this study, we collected the statistics of content accessed via different base stations, and the cosine similarity [25] was used to measure the similarity of content popularity distributions at BSs. Cosine similarity characterizes similarity by measuring the cosine of the angle between the vectors characterizing individual characteristics. The value range was $[-1, 1]$, where greater values reflect stronger similarity in the individual. The heat map shown in Fig. 3 visually demonstrates the similarity of content popularity at different base stations. Overall, there is a certain degree of similarity in the user preference at different base stations. Each base station has a certain coverage area where mobile users access different contents. Although each user has a personalized preference [11], the role of the group reduces this difference. Hence, the requested contents at different BSs share similarities. At the same time, it can be observed that there is a non-negligible difference in the content popularity at base stations. Different base stations serve different users. As observed in Fig. 2, most users have a low diversity of interests, which produces differences in content popularity among base stations with small intersections of user groups. Furthermore, similarities and differences also exist in the interest distribution of Internet content served by base stations. Hence, base stations with similar distributions of content interests can be clustered together to respond to more requests for user content by employing content-sharing. Thus, the core network traffic can be uploaded and user QoE is improved.

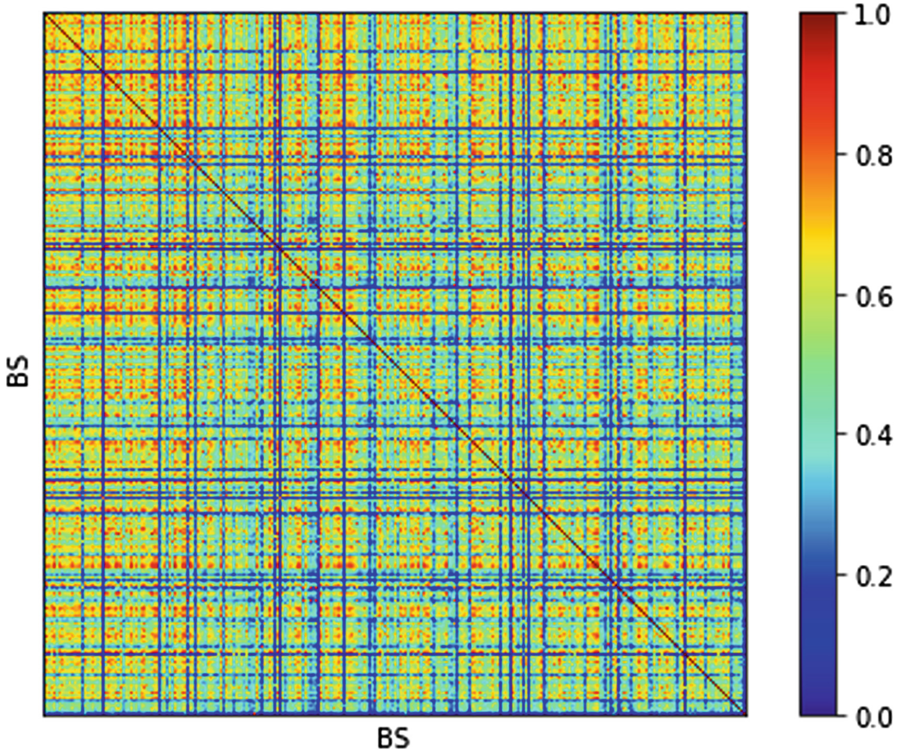

Fig. 3. Similarity of content popularity at different base stations (BSs).

Based on the above, we cache content with high popularity at BSs. Given that the content accessed is dynamic and changes over time, then how does content popularity changes with time? To answer this question, we measured the similarity of the content popularity of any two days within the 23-day observation period for the entire dataset using Jensen-Shannon (JS) divergence. The result is as depicted in Fig. 4.

The JS divergence, with range $[0, 1]$, measures the difference between the two probability distributions using the arithmetic mean of the correlation entropy of each probability distribution and the mean of the two probability distributions. When the two probability distributions are the same, the JS divergence is 0, otherwise it is 1. In this paper, 1-JS divergence is used as a measure of the degree of similarity between two probability distributions of content popularity. As observed in Fig. 4, the similarity is exhibited scores of 0.93 and higher. Note that the closer the two dates, the higher the similarity. Hence, the distribution of content popularity is stable over time, which probably results from the low diversity of user preference and the strong stickiness of Internet content. Based on this observation, we suggest that it is feasible to use online records from the previous day to guide follow-up cooperative caching.

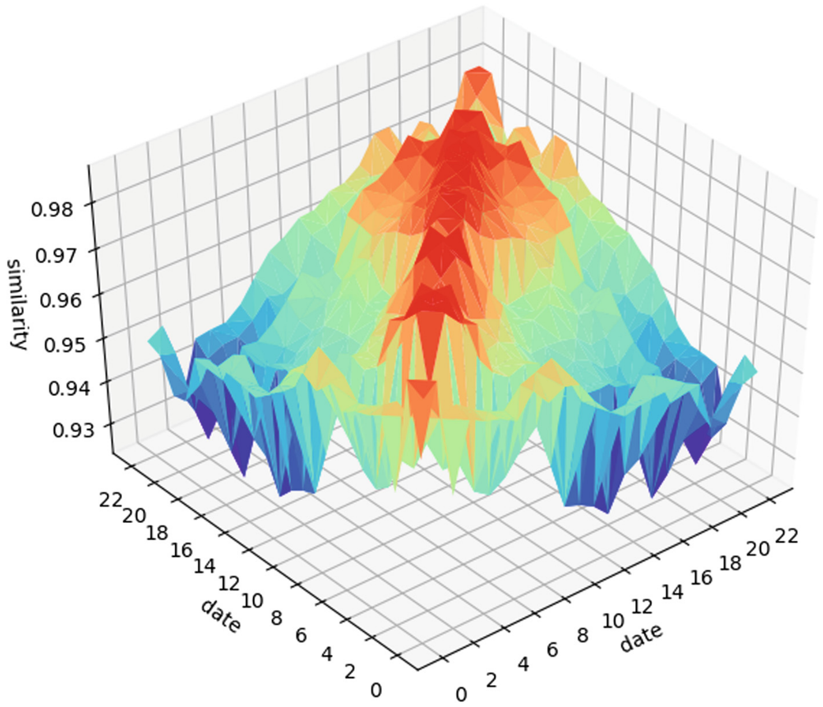

Fig. 4. Temporal stability of content popularity.

To summarize, we observe that the preferences of different users overlap and that similarities and differences exist in the interest distribution of Internet content served by base stations. As much high popularity content as possible can be cached at BSs within constraints of storage size if BSs with similar content popularity distribution can share the cached content to reduce the storage size. The stability of the content popularity distribution guarantees the feasibility of caching content with high popularity at BSs using the proposed cooperative caching and delivery algorithm.

4 System Model and Problem Analysis

4.1 System Model

The cooperative edge caching architecture focused on cooperation among BSs is illustrated in Fig. 5. Outside the mobile network operator (MNO) network, some service providers (SPs), such as YouTube and Facebook, offer content files over the Internet. Inside the MNO network, numerous BSs cover the service area. Mobile user content requests are received and served by their associated BSs, while SPs cache content onto the core network supported by the MNO. These contents are transmitted to BSs via backhaul link to satisfy the content requests of mobile users served by the BSs. Because the content requests made by different users or from different locations are inevitably intersected [5], the transmission of contents increases backhaul traffic and is frequently duplicated.

To alleviate backhaul congestion and enhance QoE, cache-enabled Cloudlets can be combined with traditional base stations distributed at the edge of the network. Each Cloudlet provides content services for mobile users in their respective regions. Data Centers (DCs) and Distributed Cloudlets use backhaul links to control information transmission and distribute content to various local caches. Each Cloudlet communicates with each other Cloudlet to transfer information and share content, reducing data traffic to and from remote DCs and improving caching capability.

Cloudlets and their neighboring BSs constitute cooperative caching domains. As shown in Fig. 5, several cooperative caching domains exist at the network edge. The caching-enabled BSs are connected and communicate with each other via X2 link [26]. To satisfy as many content requests from mobile users in the network edge as possible, the associated local BS either returns the content if it is locally available or retrieves the content from other BSs. In this way, some of the congested and costly backhaul link traffic becomes lower cost, internal traffic between BSs, reducing content access latency and increase quality of service (QoS).

Under this architecture, data content services are implemented. This paper focuses on two problems:

1. The definition of cooperative caching domains and storage of valuable contents.
2. Intercommunication between neighboring BSs in limited-capacity local caching within a cooperative domain.

In response to above issues, this paper tends to make an analysis and discussion in the following sections.

4.2 Problem Statement

Through the tensor decomposition of multi-source relational data, base station clustering with relatively consistent content access patterns and close spatial distances was achieved and base station clustering was correlated with content clustering. If the content accessed via the base station was relatively similar, and the set of users served by the base station also exhibited similarity, then these base stations were considered to have a large overlap in the content that can be cached because of the low diversity of user interests and the reduced variation in services requested by similar user sets. As content access increases in similarity, increasing portions of base station collections of user interests may be cached by each base station, without having to cache the same or similar content on each base station. Thus the base stations use surrounding base stations collaboratively. The cached content can substantially increase the heterogeneity of the content of the edge base station cache and greatly increase the utilization of cache resources, which is equivalent to expanding the cache size.

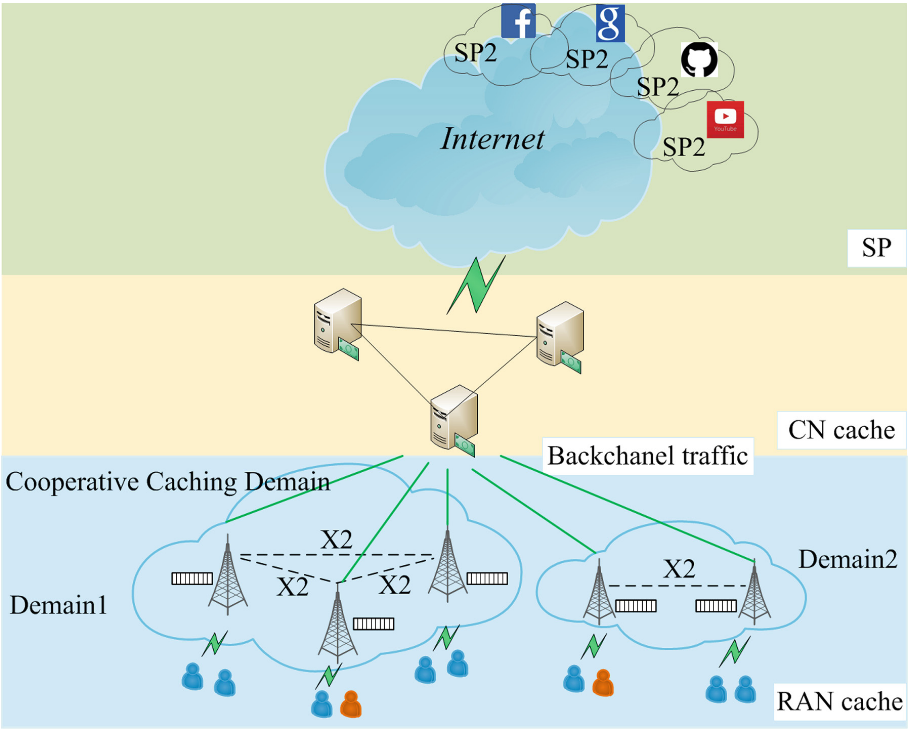

Fig. 5. Illustration of the cooperative edge caching architecture focused on collaborations among back stations (BSs).

It is worth noting that to mitigate the negative influence of content sharing between the base stations on request delay and user experience, the base stations with content sharing relationships should have shorter duration transmission time. This indicator is measured by attributes such as transmission bandwidth, transmit power, data rate, path loss index, transmission power, signal-to-interference-plus-noise ratio (SINR), channel fluctuation, and geographical distance [27]. This study defines the cost of collaboration as the effect of network content sharing among base stations on user experience. When hitting the user experience quality index and optimizing it, the negative impact of collaboration costs cannot be ignored. In this paper, the cost of content transmission between the base stations is measured by the geographical distance between base stations, as shown in (1).

$$cost_{ij} = cost(BS_i, BS_j) = f(d_{ij}) \tag{1}$$

Intuitively, to minimize the impact of content delivery on the average request latency of mobile users, the geographical distance between the base stations attempting to cooperate should also be minimized. In the case of transmission bandwidth, data rate and other parameters, the cost of cooperation can be more complex and refined. Therefore, the collaboration cost defined in this paper is universal, and can be used in applications that consider additional factors by introducing new parameters to further describe collaboration cost.

To simplify the problem complexity and highlight the content placement and transmission strategies that this study attempted to optimize, we assumed that each content file was of equal length and normalized this length to one byte. As do many papers on cooperative caching [19,28–30], we assume that it is reasonable to set the value of all file lengths to one byte because files of different lengths may be segmented into equal-length fragments that implement code-based cooperative caching [30]. Hence, the maximum amount of content that can be stored on each base station is the number of cache units.

Base station clustering is applied to form cooperative cache domains by identifying clusters of base stations with cooperative relationships. After determining the cooperation cache domain and selecting the content collections to be cached in this domain, it was necessary to determine which contents should be cached given the limited cache capacity and how the base stations should execute routing to maximize the hit rate while controlling the cooperation cost. Hence, the importance of determining the types and contents to be cached.

5 Cooperative Content Caching and Delivery Algorithm

5.1 Framework

The framework of algorithm proposed in this paper is diagrammed in Fig. 6. The framework of the method can be divided into three stages. In the first stage, the content access behavior will be extracted from the original UDRs through a 3-order tensor while simultaneously obtaining the geographical distribution of BSs. In the second stage, we use tensor decomposition to mine content access patterns. The tensor can extract and model interaction relationships between base stations, users, and content. Hence, BS clustering based on similar content access patterns that considers geographical distances informs the construction of various cooperative caching domains. In the final stage, after comprehensive consideration of the cache ratio and the content transmission cost, a distributed optimization cooperative caching and delivery algorithm is proposed.

5.2 Cooperative Caching Domain

Cooperative caching domain consists of several cooperative BSs. Base station cooperation occurs when the BSs share content by caching content at one BS and using it to satisfy content requests from mobile users served by other BSs in the same cooperative caching domain. This efficiency is based on the expected similarity of the content served by different BSs in the same caching domain.

Our previous research demonstrated that mobile user content preferences exhibit individual characteristics and stability [5]. Correspondingly, content access at BSs serving similar mobile users is expected to be similar. Therefore,

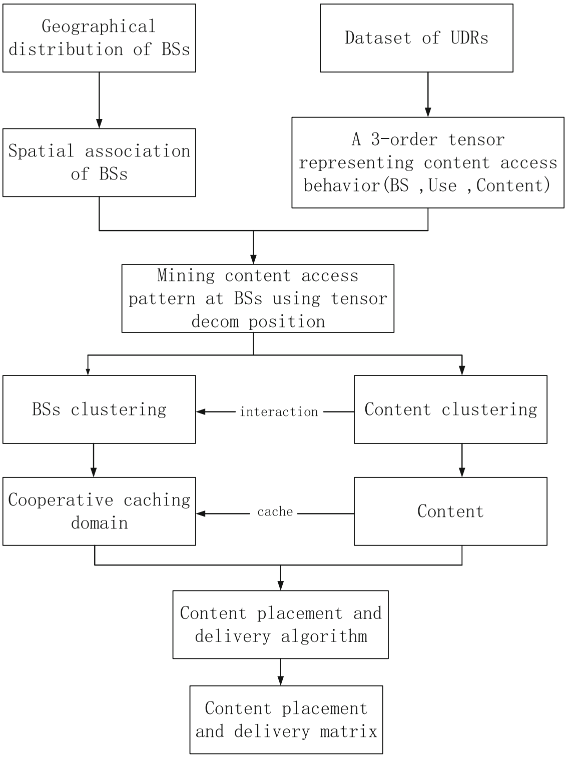
Geographical distribution of BSs
Dataset of UDRs
Spatial association of BSs
A 3-order tensor representing content access behavior(BS ,Use ,Content)
Mining content access pattern at BSs using tensor decom position
BSs clustering
interaction
Content clustering
Cooperative caching domain
cache
Content
Content placement and delivery algorithm
Content placement and delivery matrix

Fig. 6. Algorithm framework.

we assert that BSs with similar requested content and users share similar user access patterns. Thus, cooperative caching domains can be determined based on the aspects of content access and mobile users.

Like [27], we assume that caching-enabled BSs in the same cooperative domain are connected via X2 link. To reduce the transmission cost, cooperative BSs operating in the same cooperative caching domain should be close in geographical distance.

In conclusion, BSs in a single cooperative caching domain should satisfy requests for similar user access patterns within a smaller geographical distribution.

As mentioned previously, content access behavior of the base station from the service user and content must be mined and analyzed. All three dimensions, base station, user, and content, must be considered. The traditional matrix can only capture two dimensions of information [31]. In the application scenario of this paper, it is necessary to perform two-dimensional data analysis and then integrate it. In the analysis of two-dimensional data, other dimensions of information were not involved, thus introducing errors. Hence, this paper attempts to capture and analyze multiple dimensions of information simultaneously, to better examine interactions between different dimensions of information.

To discover neighbor BSs accessing similar content requested by similar users, this paper used tensor decompositions [32] to achieve pattern mining. It extended data analysis to multiple dimensions and captured multiple facets of access patterns, thus outweighing the two-dimensional representation of data [31]. Tensors captured users content access behavior at BSs while Tucker decompositions realized multi-faceted analysis for content access behavioral pattern discovery.

Using the mobile user UDR dataset, we focused on identifying patterns of content access behavior at the network edge. Let the dataset be a list of tuples (u, b, c) denoting that a mobile user via BS visits content c. We model the data as a 3-order tensor $X \in \mathbb{R}^{n_{(a)} \times n_{(b)} \times n_{(c)}}$, where $n(u)$ is the number of mobile users, $n(b)$ is the number of BSs, and $n(c)$ is the number of contents. Tensor $X(u, b, c)$ has a value of the number of existing tuples (u, b, c). Our goal is to factorize the tensor

$$X \approx G \times_{(u)} U \times_{(b)} B \times_{(c)} C \tag{2}$$

where $G \in \mathbb{R}^{R_{(a)} \times R_{(b)} \times R_{(c)}}$ is the core tensor, which encodes the behavioral patterns, i.e., the relationships among users, BSs, and content groups. The probability represented by G(r(u), r(b), r(c)) indicates the behavior of the r(u)-th user group via the r(b)-th BS group visits the r(c)-th content group, which can be expressed by simple mathematical formulae as (3)

$$\begin{aligned} &G(r^{(u)}, r^{(b)}, r^{(c)}) \approx \\ &p(user \in r^{(u)}, BS \in r^{(b)}, content \in r^{(c)} \mid (user, BS, content) happens) \end{aligned} \tag{3}$$

The user projection matrix is given as $U \in Rn(u) \times R(u)$. where U(i(u), r(u)) represents the probability that the i(u)-th user belongs to the r(u)-th group, which can be formulated as (4)

$$U(i^{(u)}, r^{(u)}) = p(user_{i^{(u)}} \mid r^{(u)}) = p(user_{i^{(u)}} \in r^{(u)}) \tag{4}$$

where $B \in Rn(b) \times R(b)$ is the BSs' projection matrix and B(i(b), r(b)) represents the probability that the i(b)-th BS belongs to the r(b)-th group, which can be formulated as (5)

$$B(i^{(b)}, r^{(b)}) = p(BS_{i_{(b)}} \mid r^{(b)}) = p(BS_{i_{(b)}} \in r^{(b)}) \tag{5}$$

where $C \in Rn(c) \times R(c)$ is the contents' projection matrix and Ct(i(c), r(c)) represents the probability that the i(c)-th content belongs to the r(c)-th group, which can be formulated as (6)

$$C(i^{(c)}, r^{(c)}) = p(content_{i_{(c)}} \mid r^{(c)}) = p(content_{i_{(c)}} \in r^{(c)}) \tag{6}$$

Note that in many applications including UDRs, the tensor is highly sparse, that is, many of the values are zero, because of the very nature of the application [32]. Commonly, elements in the tensor may be constrained by practical factors, such as distance constraint in this paper, which considers distance factors in the proposed cooperative content caching and delivery algorithm.

To solve the sparsity problem and introduce realistic constraints in tensor decompositions, as well as to guide the decomposition to identify the correct structure coinciding with features of cooperative caching domain, we introduce the BSs' geographical distance as side information and impose regularization to tackle sparsity and constraints. The regularizers can be encoded as Laplacian matrices L(b), where the (i, j)-th element represents the similarity between the i-th and j-th entities, i.e., BSs. The similarity should be inversely proportional to how far apart the BSs are located.

Referencing multi-faceted analysis method for behavioral pattern discovery proposed in [33], we incorporate the multi-faceted information and constraints into the tensor decomposition. We denote by μ(b) the weight of the BS-pattern Laplacian matrix L(b). The covariance matrix of the m-th pattern is

$$C^{(m)} = X^{(m)} X^{(m)T} + \mu^{(m)} L^{(m)} \tag{7}$$

where $X^{(m)}$ is the pattern-m matricizing of the tensor X. The projection matrices can be computed by diagonalization: they are the top r(m) eigenvectors of the covariance matrix C(m).

At the same time, benefiting from the tensor's extreme sparsity, the computational complexity of the algorithm in [33] appears as a linear relationship with the sum of the number of elements in each dimension. Additional details about multi-faceted tensor decomposition are available in [33].

To summarize, after incorporating the multi-faceted information and constraints into the tensor decomposition, clustering is executed based on BSs, content, and their interactions. The in-cluster BSs serve similar content to similar mobile users [34]. In addition, the distance constraint minimizes the in-cluster geographical distribution of BSs. As a result, BS clusters can be used to determine cooperative caching domain, while simultaneously clustering users with strong interactions with clustered BSs establishes the caching content set in a given cooperative domain.

5.3 Cooperative Content Caching and Delivery Method

After determining the cooperative caching domain and its content set, we should consider efficient content placement and delivery algorithm with limited caching capacity to maximize cache hit ratios and reduce cooperative cost. It is crucial to decide which files should be cached and where to cache them.

Given N BSs composing R collaboration caching domains, where the r-th collaboration caching domain contains N_r BSs. The frequency of an interaction involving BS_j in one cooperative caching domain is ω_j, considered to be the probability of BS_j belonging to this cooperative caching domain. The content set associated strongly with the r-th cooperative caching domain is $\{f_1, f_2, ...F_n, ...f_{F_r}\}$,where $F_r = N_r * cachesize$, $cachesize$ denotes the size of the space available for BS caching. The number of files involved is F, the popularity of f_n at BS_j is $(p_j)^n$.

For convenience, we assumed that each file has the same file length and normalized to 1 byte. This is reasonable because files of different length can be divided into groups of the same length. Thus, the maximum number of files each BS can store is also the cache size. Constraint to the feature of dataset, we measure transmission cost in terms of geographical distance, namely, $cost_{ij} = cost(BS_i, BS_j) = f(d_{ij})$. In real-world experience, the cooperative cost is related to many factors such as transmission bandwidth, data rate, path loss index, signal-to-interference-plus-noise (SINR) ratio, and geographical distance. With these factors obtained, the cooperative cost realizes greater complexity and sophistication. The cost we defined here has universality. In applications considering more factors, we can incorporate the additional factors in the definition of cooperative cost.

To achieve a distributed caching mechanism aimed at maximizing the cache hit ratio and reducing the cooperative cost, we must determine which files are to be cached and where to cache them in every cooperative caching domain. Meanwhile, we must also determine which BSs can satisfy the content requests that cannot be served locally. That is, we must solve the problems of what to cache, where to cache, and how to cooperate, which is realized by the proposed caching placement and delivery algorithm.

Content placement matrix $(x_j^f)_{N\times F}$, where $x_j^f \in 0, 1$ denotes whether or not f is cached at BS_j.

Content delivery matrix $\delta = {\delta_{jk}^f}_{N\times N\times F}$, where $\delta_{jk}^f \in \{0,1\}$ denotes that whether BS_j will ask BS_k to transfer the file f which is stored not in BS_j but in BS_k and is requested by users serviced by BS_j.

$$\max_{x_j^f,\delta_{jk}^f} \sum_{r=1}^{R}(\sum_{j=1}^{Nr}\omega_j \sum_{f=1}^{Tr} p_j^f(x_j^f + \sum_{k=1}^{Nr}\delta_{jk}^f)) \tag{8a}$$

$$\min_{\delta_{jk}^f} \sum_{r=1}^{R}(\sum_{j=1}^{Nr}\sum_{k=1}^{Nr}\sum_{f=1}^{Tr}\delta_{jk}^f \cdot p_j^f \cdot e^{d_{jk}}) \tag{8b}$$

$$s.t. \qquad x_j^f + \sum_{k=1}^{Ni}\delta_{jk}^f \le 1 \qquad \forall i, \forall f \tag{8c}$$

$$delta_{jk}^f \le X_k^f \qquad \forall i, \forall k, \forall f \tag{8d}$$

$$\sum_{f=1}^{F} x_j^f \le cachesize \qquad \forall j \tag{8e}$$

$$\delta_{jj}^f = 0 \qquad \forall j, \forall f \tag{8f}$$

$$x_j^f \in 0, 1, \delta_{jk}^f \in 0, 1 \qquad \forall i, \forall k, \forall f \tag{8g}$$

- Our objective is measured in terms of both the hit ratio and cost of the information exchange between BSs incurred by the users of each cooperative caching domain. The objective is to maximize the hit ratio for serving users, while minimizing the cooperation cost, as shown in (8a) and (8b).

- The constraints in (8c) denote the cooperation among BSs and guarantee that any content request will not be routed to other BSs if the content is locally available.
- Constraint (8d) states that a content can be fetched from the BS only if the BS caches that content.
- The constraints in (8e) enforce resource limits at each BS, stating that all the items cached at a BS cannot exceed its cache capacity.
- Constraint (8f) guarantees the nonexistence of the content request transferring to local BS.
- Constraint (8g) defines this problem as 0–1 integer programming.

The problems expressed in (8a) and (8b) constitute a multi-objective optimization problem. The value range and dimension of the hit ratio and cooperative cost are not uniform. Therefore, we first set the min-max standardization of the collaboration cost to be in the range of $[0, 1]$. Then we used the linear weighted method to integrate the two optimization goals into a single optimization formula, given as (9).

$$min_{x_j^f, \delta_{jk}^f} \sum_{r=1}^{R} (S(\sum_{j=1}^{Nr} \sum_{k=1}^{Nr} \sum_{f=1}^{Tr} \delta_{jk}^f \cdot p_j^f \cdot e^{d_{jk}})) \tag{9}$$

The standardization of the collaboration cost is shown in (10).

$$S(x) = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{10}$$

In the application scenario of this article, as expressed in (10), $\min(x) = 0$, $\max_{\substack{j,k \in Br \\ f \in Cr}} \sum_{j=1}^{Nr} \sum_{k=1}^{Nr} \sum_{f=1}^{Tr} p_j^f \cdot e^{d_{jk}}$.

When the linear weighting method is used to solve multi-objective optimization problems, the weighting coefficients must be carefully selected according to the problem scenario. However, in this paper, the coordination cost is normalized so that it is within the range of fluctuations that are equivalent to the hit rate, effectively avoiding problem-dependent weight coefficient selections that could introduce excessive subjectivity.

To obtain the optimal solution under the constraint conditions described by (8c) through (8f) for the optimization goals declared in (8a) and (8b), this paper uses Lagrangian duality to convert the original problem into a dual problem and then apply the solution to the dual problem to obtain the solution to the original problem. By introducing the Lagrangian multiplier, the optimization problem with d variables and k constraints can be transformed into an unconstrained optimization problem with $\mathrm{d} + \mathrm{k}$ variables.

First, we investigate the feasibility of solving the dual problem instead of solving the original problem. If the decision variable is a continuous value in the range $[0, 1]$, then the optimization objectives of (8a) and (8b) and the inequality constraints given in (8c) through (8e) are all linear functions related to the decision variable, belonging to the convex function. One function, the equality constraint (8f), belongs to the affine function, and the presence of a certain value of the decision variable makes the inequality constraint strictly feasible, so there is a precondition for solving the original problem by solving the dual problem [32]. To take advantage of the computational simplicity of the dual problem, we relaxed the original problem into a generalized linear program (LP)-type problem that does not contain integer constraints. Then, we used the Lagrangian duality to convert it into a dual problem, as delineated in (11).

$$\begin{aligned} L(x,\delta,\alpha,\beta,\gamma,\omega,\mu,\eta,\nu,\sigma) = \sum_{r=1}^{R}(S(\sum_{j=1}^{Nr}\sum_{k=1}^{Nr}\sum_{f=1}^{Tr}\delta_{jk}^{f}\cdot p_{j}^{f}\cdot e^{d_{jk}}) \\ -\sum_{j=1}^{Nr}\omega_{j}\sum_{f=1}^{T}p_{j}^{f}(x_{j}^{f}+\sum_{k=1}^{Nr}\delta_{jk}^{f} \\ +\alpha_{j}^{k}(x_{j}^{k}+\sum_{k=1}^{Nr}\delta_{jf}^{f}-1)+\beta_{jk}^{f}(\delta_{jk}^{f}-x_{k}^{f}) \\ +\gamma_{j}(\sum_{f=1}^{Tr}s_{j}^{f}-cachesize)-\omega_{j}^{f}x_{j}^{f}-\mu_{jk}^{f}\delta_{jk}^{f} \\ +\eta_{j}^{f}(x_{j}^{f}-1)+\nu_{jk}^{f}(1-\delta_{jk}^{f}+\sigma_{j}^{f}\delta_{jj}^{f})) \end{aligned} \tag{11}$$

where are Lagrange multipliers and non-negative. Then, in (12), we solve the dual problem.

$$\max_{\substack{\alpha,\beta,\gamma,\omega,\mu,\eta,\nu,\sigma \\ \alpha,\beta,\gamma,\omega,\mu,\eta,\nu\geq 0}} \min_{x,\delta} L(x,\delta,\alpha,\beta,\gamma,\omega,\mu,\eta,\nu,\sigma) \tag{12}$$

Solving the dual problem can obtain an analytical solution without having to iterate to obtain the numerical solution, which can greatly reduce the solution complexity. Finally, the connection between the solution to the original problem x^*, δ^* and the optimal solution to the dual problem $\alpha^*, \beta^*, \gamma^*, \omega^*, \mu^*, \eta^*, \nu^*, \sigma^*$ is established through the Karush-Kuhn-Tucker (KKT) condition [35].

Because of the relaxation of the original problem, in using the sequential minimal optimization (SMO) algorithm to solve, we must integerize the non-integer solution. Assume that the non-integer values between $[0, \mathrm{a}]$ and $[\mathrm{a}, 1]$ $(0 \leq a \leq 1)$ are integerized to 0 and 1 respectively, and then driven by the optimization goal expressed in (11). A grid search is used to determine the rounding decision boundaries of content placement decision variables and content delivery decision variables. Based on this, the content placement matrix and routing matrix can be obtained. While controlling the collaboration cost, the hit rate of the cache resources is high, thus satisfying more user content requests at the edge of the network.

6 Performance Evaluation

6.1 Simulation Setup

We used the UDR dataset introduced in Sect. 3 to evaluate the proposed algorithm. Because of limited cache space and the power-law distributions of access popularity, most contents are rarely accessed. Hence, we preprocessed the data to remove contents with small popularity values. To reduce the complexity of tensor decomposition and the distributed caching mechanism, contents with large popularity values were selected for caching. In this paper, the first three days of UDRs were used as a training set and the test set comprised the remaining twenty days of data. The tensor was obtained based on the training set. The proportion of non-zero elements in the tensor was as low as 5.194e−09. After preprocessing the dataset, the proportion of non-zero elements in the tensor was as low as 2.17943605808e−05. This demonstrates that the tensor representing the interactions among users, BSs, and contents derived from UDRs is extremely sparse and well-suited to storage and decomposition.

In this experience, we obtained a BS distance matrix based on the geographical position information of BSs. The user access patterns at the network edge were calculated using the Tucker decomposition subject to distance constraints. Per the objective of (7), we have designed a content placement and delivery algorithm. We evaluated the performance of the proposed algorithm using the test set.

6.2 Performance Results

We considered the following four content caching schemes:

1. User_Cluster: Users are clustered and BSs providing contents are allocated to maximize the hit ratio [17].
2. BS_Cluster_Geo: Base stations are clustered based on geographical position [11].
3. BS_Cluster_ContentModePattern: Based on user access patterns at the network edge, the purpose is to identify cooperative caching domains and the content contained in each domain without considering a distance constraint.
4. BS_Cluster_ContentModePattern_Geo: Based on user access patterns at the network edge and a distance constraint, the proposed algorithm attempts to identify cooperative caching domains and their respective contents.

We compare the efficiency of the four content caching schemes using the UDR dataset. After determining the cooperative domain, all four schemes designed content placement and delivery algorithms that were evaluated based on their performance on the test set as defined by the content distribution mechanism in (7). Specifically, we compared performance using three metrics, the hit ratio, cooperative cost, and the content utilization ratio, and then considered stability over time.

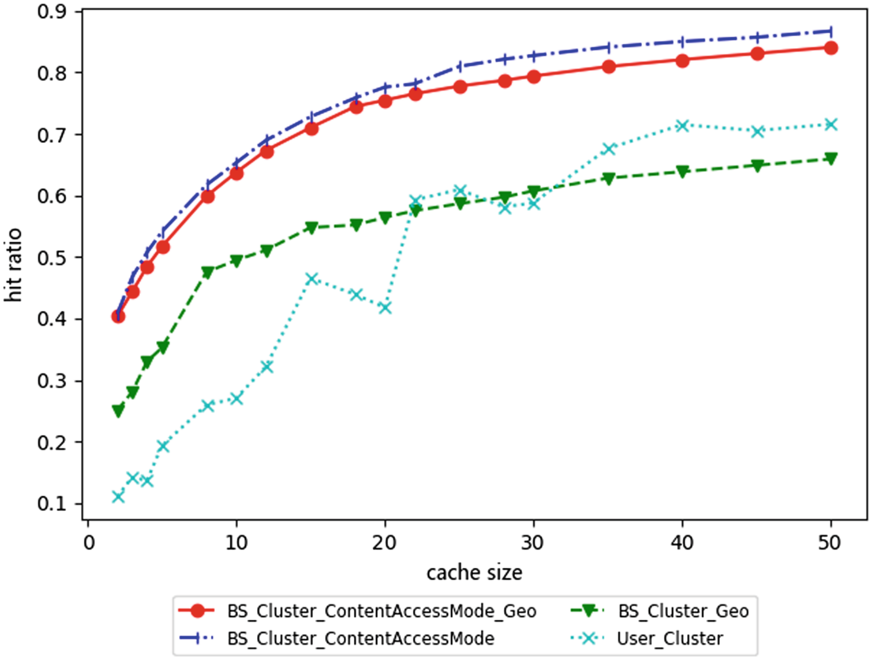

Fig. 7. Cache hit ratio versus cache size.

Hit Ratio. First, we investigated the hit ratio. Figure 7 depicts the total cache hit ratio of UDRs collected during the 20-day period covered by the test set, for different cache sizes C, we considered BS cache sizes of two to fifty contents. The simulations suggest that the cache ratio of all the cooperative caching mechanisms is positively related to the cache size. With the increase of cache size, the impact on the hit ratio slows down. Overall, in terms of satisfying user content requests, cooperative caching that considers user access patterns achieves a hit ratio increase of approximately 25% compared to cooperative caching based on BS or user clustering. However, enforcing distance constraints resulted in a slight decrease in the hit ratio. This demonstrates the superiority of the proposed method in satisfying user content requests.

Furthermore, we note that even when the cache storage size was small, many user content requests were satisfied. For example, when cache size was 10,0 63% of user content requests were satisfied. This demonstrates that less popular content can cover most user interests, which is consistent with the power-law characteristic of content access behavior [4].

Cooperative Cost. The definition of cooperative cost is in part D of Method. The cooperative cost which is for all UDRs in the test set in Fig. 8 show a trend of increasing first and decreasing then with the increase of cache size. When the cooperative size is small, the most popular contents are cached and BSs can achieve more cooperative probability. With the increase of cache size and constraints of cooperative cost, the most popular contents are cached at BSs locally and the content transfer is decreased gradually.

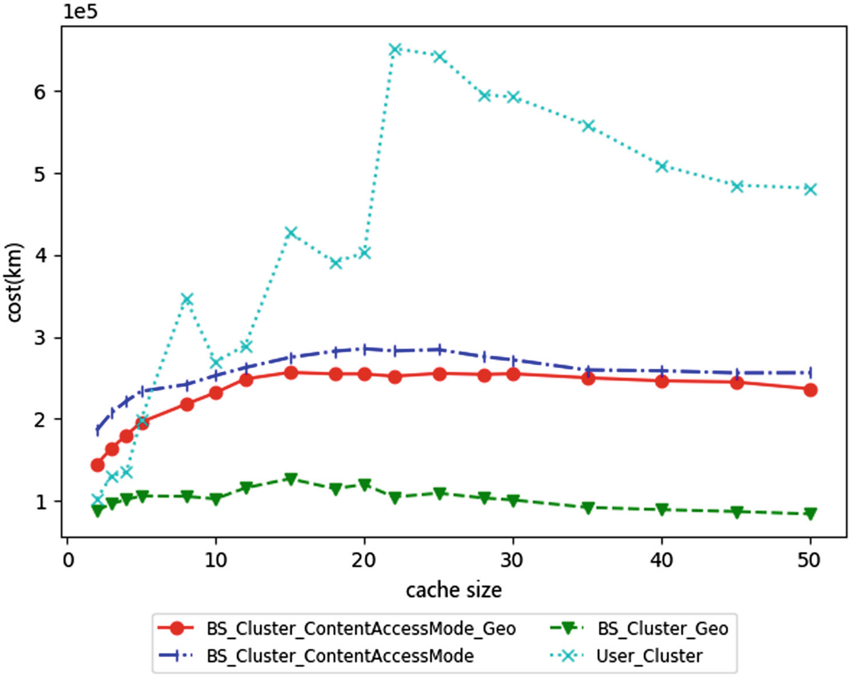

Fig. 8. Cooperative cost versus cache size.

The transmission cost of cooperative caching that considers user access patterns is higher than that based on BS clustering and lower than that based on user clustering. When setting the file size to 1 byte, cooperative caching based on the geographical positions of BSs has a lower transmission cost, but this is paired with a lower hit ratio, as observed in Fig. 7. Cooperative cost increases for cooperative caching based on user clustering, as this approach does not consider the geographical distribution of users, and thus clusters users with varied geographical distances and similar interests. Although the proposed method includes a distance constraint that slightly reduces the hit ratio, the constraint also reduces the cooperative cost.

Content Utilization Ratio. The hit ratio indicates the probability of responses to user content requests, while how many the cached files users can access is referred to as the content utilization ratio. Figure 9 depicts the content utilization ratio achieved by each algorithm with various cache sizes. As cache size increases, the utilization ratio of cached content correspondingly decreases.

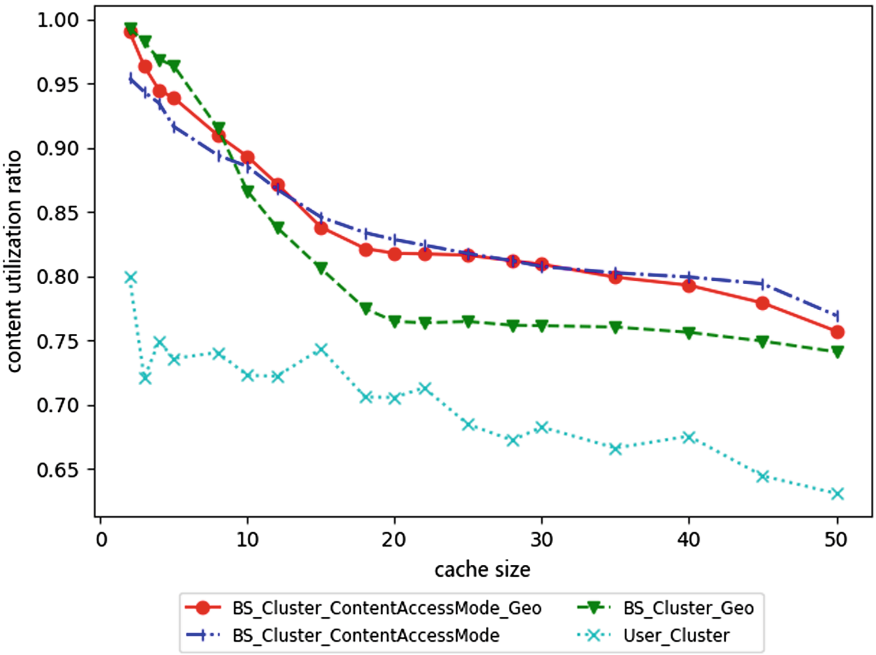

Fig. 9. Caching utilization ratio versus cache size.

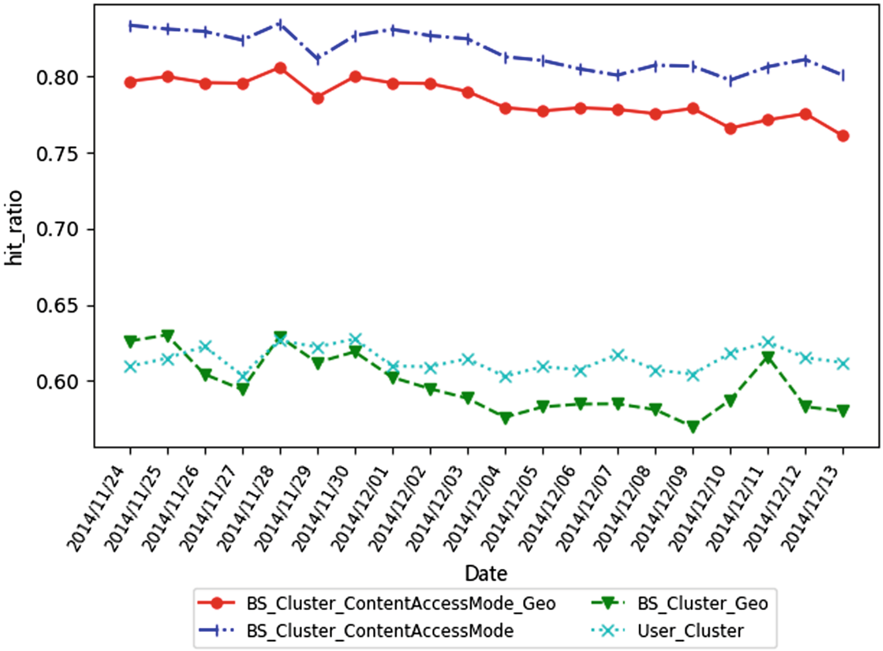

Fig. 10. Stability of hit ratio over time.

Cooperative caching based on user access patterns achieves the highest utilization ratio. As previously demonstrated, consideration of distance constraints balanced the hit ratio against cooperative cost. When the cache size was small, the distance constraint lowered transmission cost while raising content utilization ratio. As cache size increased, so did the diversity of the contents, which then decreased the possibility of sharing contents. We highlight the inference effect of the distance constraint on the similarity of BSs and the corresponding decrease in content utilization ratio.

When the cached files are few, the corresponding utilization ratio and the hit ratio are both high. This indicates that these files are the most frequently requested and that the cache space is fully utilized [4].

Based on our analysis, the proposed method improved the cache hit ratio while balancing the hit ratio and the cooperative cost using the distance constraint. Although the distance constraint slightly decreased the hit ratio and content utilization ratio, these losses were offset by benefit of decreased cooperative cost.

Stability over Time. We use the content placement matrix and delivery matrix in derived from the test set to evaluate the stability over time, as depicted by Figs. 10 and 11. Note that cooperative caching based on user access patterns has

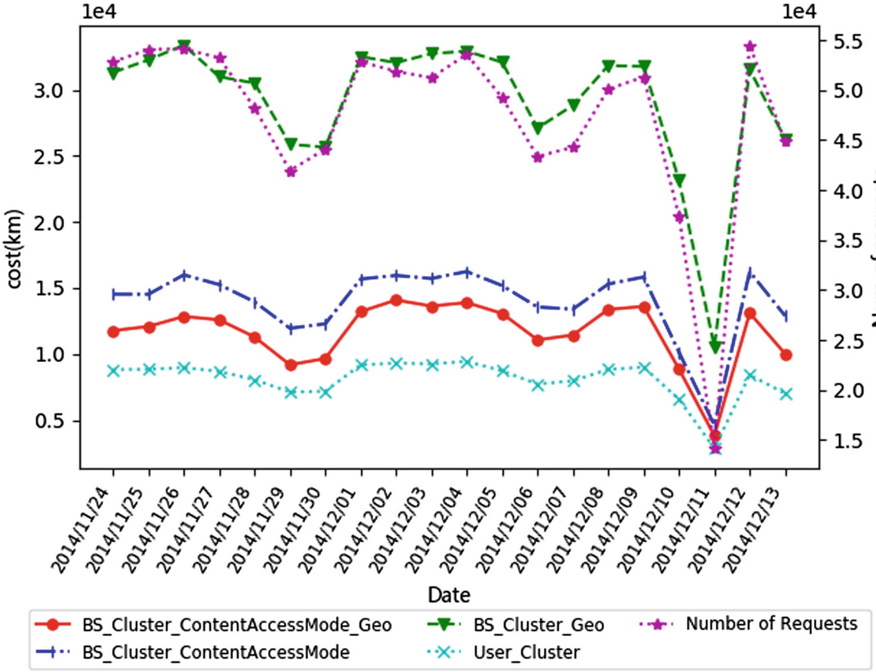

Fig. 11. Stability of cooperation cost over time.

a higher hit ratio, which is consistent with observation regarding Fig. 7. In terms of stability over time, cooperative caching based on user access patterns was superior to that based on user clustering, while slight fluctuations were observed in cooperative caching based on BS clustering.

From the results, we establish that the proposed method improves user QoE and is stable over time. The content placement caching algorithm also guarantees the long-term popularity of cached contents. Meanwhile, the long-term stabilization effect can also be observed using the three days of training data, underscoring the stability of the proposed caching algorithm even when applied to small datasets. This attribute indicates that the proposed method could be applied to mobile big data scenarios.

7 Conclusion

This paper proposed cooperative caching based on user access patterns at the network edge. Tensor decompositions of multi-aspect data were applied. We find that BSs with close geographical distances accessed similar content at the request of similar users. There was a greater demand for content sharing among these BSs and the costs were not high, thus hey constituted cooperative caching domains. Furthermore, we designed a content placement algorithm that simultaneously considered the hit ratio and the cooperative cost. In every cooperative caching domain, the contents that frequently interacted with the BSs in the domain were placed and shared accordingly.

We evaluate the performance using a real UDR dataset. The results show that the method proposed in this paper can improve the caching hit ratio and the content cache utilization ratio while moderating cooperative cost and maintaining stability overlong time. The overall performance was superior, and thus able to meet user content access demand and be applied to mobile data.

Acknowledgment. This work was supported in part by the National Natural Science Foundation of China (NSFC), Grant No. 61371126. This work was financially supported by self-determined research funds of CCNU from the colleges basic research and operation of MOE (CCNU17QN0011) and specific funding for education science research by self-determined research funds of CCNU from the colleges basic research and operation of MOE (ccnu16JYKX019).

References

1. Ahmed, A., Ahmed, E.: A survey on mobile edge computing. In: International Conference on Intelligent Systems and Control (2016)
2. Wang, S., Zhang, X., Zhang, Y., Wang, L., Yang, J., Wang, W.: A survey on mobile edge networks: convergence of computing, caching and communications. IEEE Access **5**(99), 6757–6779 (2017)
3. Li, X., Wang, X., Li, K., Leung, V.: CaaS: caching as a service for 5G networks. IEEE Access **5**, 5982–5993 (2017)

4. Wang, C.X., et al.: Cellular architecture and key technologies for 5G wireless communication networks. J. Chongqing Univ. Posts Telecommun. **52**(2), 122–130 (2014)
5. Zhou, C., Jiang, H., Chen, Y., Wu, L., Yi, S.: User interest acquisition by adding home and work related contexts on mobile big data analysis. In: Computer Communications Workshops (2016)
6. Zhou, C., Jiang, H., Chen, Y., Wu, J., Zhou, J., Wu, Y.: TCB: a feature transformation method based central behavior for user interest prediction on mobile big data. Int. J. Distrib. Sens. Netw. **12**(9) (2016)
7. Agiwal, M., Roy, A., Saxena, N.: Next generation 5G wireless networks: a comprehensive survey. IEEE Commun. Surv. Tutor. **18**(3), 1617–1655 (2017)
8. Bastug, E., Bennis, M., Debbah, M.: Living on the edge: the role of proactive caching in 5G wireless networks. IEEE Commun. Mag. **52**(8), 82–89 (2014)
9. Ramanan, B.A., Drabeck, L.M., Haner, M., Nithi, N.: Cacheability analysis of http traffic in an operational LTE network, pp. 1–8 (2013)
10. Chen, Z., Lee, J., Quek, T.Q.S., Kountouris, M.: Cluster-centric cache utilization design in cooperative small cell networks. In: IEEE International Conference on Communications (2016)
11. Chen, Z., Lee, J., Quek, T.Q.S., Kountouris, M.: Cooperative caching and transmission design in cluster-centric small cell networks. IEEE Trans. Wirel. Commun. **16**(5), 3401–3415 (2017)
12. Wang, Z., Ng, D.W.K., Wong, V.W.S., Schober, R.: Transmit beamforming for QoE improvement in C-RAN with mobile virtual network operators. In: IEEE International Conference on Communications, pp. 1–6 (2016)
13. Hu, H., Wen, Y., Niyato, D.: Spectrum allocation and bitrate adjustment for mobile social video sharing: a potential game with online QoS learning approach. IEEE J. Sel. Areas Commun. **35**(4), 935–948 (2017)
14. Li, X., Wang, X., Xiao, S., Leung, V.C.M.: Delay performance analysis of cooperative cell caching in future mobile networks. In: IEEE International Conference on Communications, pp. 5652–5657 (2015)
15. Fan, S., Zheng, J., Xiao, J.: A clustering-based downlink resource allocation algorithm for small cell networks. In: International Conference on Wireless Communications & Signal Processing, pp. 1–5 (2015)
16. Yan, H., Gao, D., Su, W., Foh, C.H., Zhang, H., Vasilakos, A.V.: Caching strategy based on hierarchical cluster for named data networking. IEEE Access **5**, 8433–8443 (2017)
17. Elbamby, M.S., Bennis, M., Saad, W., Latva-Aho, M.: Content-aware user clustering and caching in wireless small cell networks. In: International Symposium on Wireless Communications Systems, pp. 945–949 (2014)
18. Hajri, S.E., Assaad, M.: Caching improvement using adaptive user clustering. In: IEEE International Workshop on Signal Processing Advances in Wireless Communications, pp. 1–5 (2016)
19. Poularakis, K., Iosifidis, G., Tassiulas, L.: Approximation caching and routing algorithms for massive mobile data delivery. In: Global Communications Conference, pp. 3534–3539 (2014)
20. Yu, R., et al.: Enhancing software-defined ran with collaborative caching and scalable video coding. In: ICC 2016–2016 IEEE International Conference on Communications, pp. 1–6 (2016)
21. Jiang, W., Feng, G., Qin, S.: Optimal cooperative content caching and delivery policy for heterogeneous cellular networks. IEEE Trans. Mob. Comput. **16**(5), 1382–1393 (2017)

22. Borst, S., Gupta, V., Walid, A.: Distributed caching algorithms for content distribution networks. In: Conference on Information Communications, pp. 1478–1486 (2010)
23. Bao, J., Zheng, Y., Wilkie, D., Mokbel, M.: Recommendations in location-based social networks: a survey. Geoinformatica **19**(3), 525–565 (2015)
24. Ren, X.Y., Song, M.N., De Song, J.: Context-aware point-of-interest recommendation in location-based social networks. Chin. J. Comput. (2017)
25. Sidorov, G., Gelbukh, A., Gómezadorno, H., Pinto, D.: Soft similarity and soft cosine measure: similarity of features in vector space model. Computación Y Sistemas **18**(3), 491–504 (2014)
26. Ranaweera, C., Wong, E., Lim, C., Nirmalathas, A.: Next generation optical-wireless converged network architectures. IEEE Netw. **26**(2), 22–27 (2012)
27. Wang, S., Zhang, X., Yang, K., Wang, L., Wang, W.: Distributed edge caching scheme considering the tradeoff between the diversity and redundancy of cached content. In: IEEE/CIC International Conference on Communications in China, pp. 1–5 (2016)
28. Sermpezis, P., Spyropoulos, T., Vigneri, L., Giannakas, T.: Femto-caching with soft cache hits: improving performance with related content recommendation. In: GLOBECOM 2017–2017 IEEE Global Communications Conference, pp. 1–7 (2018)
29. Golrezaei, N., Shanmugam, K., Dimakis, A.G., Molisch, A.F.: Femtocaching: wireless video content delivery through distributed caching helpers. In: IEEE INFOCOM, pp. 1107–1115 (2013)
30. Borst, S.C.: Distributed caching algorithms for content distribution networks, **54**(1), 1–9 (2015)
31. Acar, E., Çamtepe, S.A., Krishnamoorthy, M.S., Yener, B.: Modeling and multiway analysis of chatroom tensors. In: IEEE International Conference on Intelligence and Security Informatics, pp. 256–268 (2005)
32. Papalexakis, E.E., Faloutsos, C., Sidiropoulos, N.D.: Tensors for data mining and data fusion: models, applications, and scalable algorithms. ACM (2016)
33. Jiang, M., Cui, P., Wang, F., Xu, X., Zhu, W., Yang, S.: FEMA: flexible evolutionary multi-faceted analysis for dynamic behavioral pattern discovery, pp. 1186–1195 (2014)
34. Schein, A., Zhou, M., Blei, D.M., Wallach, H.: Bayesian Poisson tucker decomposition for learning the structure of international relations (2016)
35. Joachims, T.: Training linear SVMs in linear time. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 217–226 (2006)