工程科学学报
Chinese Journal of Engineering

用户属性感知的移动社交网络边缘缓存机制

杨静 武佳 李红霞

User-aware edge-caching mechanism for mobile social network

YANG Jing, WU Jia, LI Hong-xia

引用本文:
杨静, 武佳, 李红霞. 用户属性感知的移动社交网络边缘缓存机制[J]. 工程科学学报, 2020, 42(7): 930-938. doi: 10.13374/j.issn2095-9389.2019.07.12.001
YANG Jing, WU Jia, LI Hong-xia. User-aware edge-caching mechanism for mobile social network[J]. *Chinese Journal of Engineering*, 2020, 42(7): 930-938. doi: 10.13374/j.issn2095-9389.2019.07.12.001

在线阅读 View online: https://doi.org/10.13374/j.issn2095-9389.2019.07.12.001

工程科学学报, 第 42 卷, 第 7 期: 930−938, 2020 年 7 月
Chinese Journal of Engineering, Vol. 42, No. 7: 930−938, July 2020
https://doi.org/10.13374/j.issn2095-9389.2019.07.12.001; http://cje.ustb.edu.cn

用户属性感知的移动社交网络边缘缓存机制

杨　静[1,2,3)], 武　佳[1,2,3)]✉, 李红霞[4)]

1) 重庆邮电大学通信与信息工程学院, 重庆 400065　2) 重庆高校市级光通信与网络重点实验室, 重庆 400065　3) 泛在感知与互联重庆市重点实验室, 重庆 400065　4) 中国联合网络通信有限公司重庆市分公司, 重庆 401123
✉通信作者, E-mail: 1309431264@qq.com

摘　要　针对数据流量爆发式增长所引发的网络拥塞、用户体验质量恶化等问题, 提出一种用户属性感知的边缘缓存机制. 首先, 利用隐语义模型获知用户对各类内容的兴趣度, 进而估计本地流行内容, 然后微基站将预测的本地流行内容协作缓存, 并根据用户偏好的变化, 将之实时更新. 为进一步减少传输时延, 根据用户偏好构建兴趣社区, 在兴趣社区中基于用户的缓存意愿和缓存能力, 选择合适的缓存用户缓存目标内容并分享给普通用户. 结果表明, 所提机制性能优于随机缓存及最流行内容缓存算法, 在提高缓存命中率、降低传输时延的同时, 增强了用户体验质量.

关键词　移动社交网络; 边缘缓存; 流行度预测; 隐语义模型; 社会属性
分类号　TN929.5

User-aware edge-caching mechanism for mobile social network

YANG Jing[1,2,3)], *WU Jia*[1,2,3)]✉, *LI Hong-xia*[4)]

1) School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
2) Chongqing Key Laboratory of Optical Communication and Networks, Chongqing 400065, China
3) Chongqing Key Laboratory of Ubiquitous Sensing and Networking, Chongqing 400065, China
4) Chongqing Branch of China Unicom, Chongqing 401123, China
✉ Corresponding author, E-mail: 1309431264@qq.com

ABSTRACT　With the rapid growth in the number of intelligent terminal devices and wireless multimedia applications, mobile communication traffic has exploded. The latest report from Cisco Visual Networking Index (CVNI) indicates that by 2022, global mobile data traffic will have grown to three times that in 2017, which will exert tremendous pressure on the backhaul link. One key approach to solve this problem is to cache popular content at the edges (base stations and mobile devices) and then bring the requested content from the edges close to the user, instead of obtaining the requested content from the content server through backhaul networks. Thus, by obtaining the required content of mobile users locally, edge caching can effectively improve network performance and reduce the pressure on the backhaul link. However, owing to the limited storage capacity of the edge nodes and the diversification of user requirements, the edge nodes can neither cache all the content in the content server nor randomly cache the content. To solve these problems, an edge-caching mechanism based on user-awareness was proposed. First, using an implicit semantic model, we predicted popular content in a macro cell in terms of the users' interests. Small base stations within identical macro cells cache data cooperatively, which update local popular content based on the dynamic content preference of users. To further reduce the delay in content delivery, we helped users to ascertain their top communities of interest based on their content preferences. At the same time, the most appropriate user equipment (UE) is selected considering the caching willingness and caching ability to cache data for other UEs in identical communities of interest. Results show that the proposed mechanism outperforms the random cache approach and the most popular content-caching

收稿日期: 2019−07−12
基金项目: 国家自然科学基金资助项目 (61771082, 61871062); 重庆市高校创新团队建设计划资助项目 (CXTDX201601020)

algorithm; it improves the cache hit rate and reduces the transmission delay while enhancing the quality of user experience.

KEY WORDS　mobile social network; edge cache; popularity prediction; implicit semantic model; social attribute

随着智能终端设备数量以及无线多媒体应用的快速增长，使得移动通信流量爆发式增长，根据美国思科公司的思科视觉网络指数预测，到2022年全球移动数据流量将增长至2017年的3倍[1]，这将给回程链路造成巨大压力．移动社交网络是一个承载在无线通信网络上的社交网络，不仅具有无线通信网络中的移动特性也具有社交网络的社会特性[2–3]，基于其移动与社会特性，用户可通过相遇机会进行内容分享，从而减少从内容服务器中获取内容的次数，减轻回程链路压力，同时，在移动社交网络中，微基站通常被部署在靠近用户终端，具有较强的储存和计算能力，相比于距离较远的云，将内容缓存在微基站可减少内容传输延迟，提高用户体验质量．

然而，由于边缘节点（终端用户、微基站）的存储容量有限以及用户需求的多样化[4]，使得边缘节点既不能缓存内容服务器中所有的内容，也不能在存储容量的限制下，随机缓存内容，因此缓存内容的选择对提高边缘缓存性能至关重要，同时，用户的移动性是移动社交网络中实现内容分享的重要特性，但也可能因为移动性使缓存用户和请求用户不在终端直接通信技术（D2D）通信范围内，最终导致内容分享失败，因此，为提高缓存命中率、减少传输时延，选择合适的缓存用户对提高边缘缓存性能至关重要．

综上所述，缓存内容选择以及缓存用户选择是提高网络性能、保证用户体验质量的关键问题．在文献中，这两个问题通常被独立的研究．对于缓存内容选择的问题，通常是选择流行的内容，为此需要对内容流行度进行预测，文献[5]使用聚类分析将内容划分为不同类型，利用自回归积分移动平均模型预测不同类型的内容流行度；文献[6]使用印度自助餐过程来分析内容选择问题，进而预测内容流行度；文献[7]基于用户的上下文，使用回声状态网络来预测用户的内容请求分布，进而预测出内容流行度．文献[5]～[7]虽然在一定程度上提升了网络性能，但都没有考虑缓存用户的选择问题．对于缓存用户选择的问题，文献[8]提出一种基于用户重要性的缓存策略，基于用户的物理属性与社会行为选择排名较高的用户来缓存内容；文献[9]提出一种分布式缓存用户选择方案，基于概率论模型来选择高速缓存用户，以防止内容的重传；文献[10]提出一种缓存用户选择方案，基于用户设备的剩余功率以及内容的最低传输功率来选择分数高的用户作为缓存用户．然而，文献[8]～[10]主要聚焦在缓存用户的选择问题，没有对内容流行度进行预测．目前，很少有文献将内容流行度的预测与缓存用户的选择结合在一起．

为此，本文提出一种用户属性感知的边缘缓存机制（User-aware edge caching mechanism, UAEC），为提高缓存命中率，减少传输时延，首先，考虑用户的偏好，预测出本地流行内容，并通过微基站将之协作缓存，同时根据用户偏好的变化将之实时更新，然后，为进一步减少传输时延，基于用户偏好构建兴趣社区，在每个兴趣社区中，考虑用户的缓存意愿及缓存能力选择合适的缓存用户缓存目标内容并分享给普通用户．

1　系统模型

考虑一个宏基站小区内的边缘缓存问题，如图1所示，该模型由内容服务器、宏基站、微基站和终端用户组成．内容服务器负责存储用户所需的所有内容$F=\{f_1,f_2,\cdots,f_N\}$，并假设每个内容的大小相同为d，宏基站负责管理整个小区内的缓存资源、计算资源和通信资源，为充分利用缓存资源，减少缓存冗余，微基站负责将本地流行内容协作缓存，并通过蜂窝链路发送到内容请求者，终端用户分为缓存用户和普通用户，普通用户可从缓存用户中获取所需内容，考虑到用户的自私性可能不愿意缓存其不感兴趣的内容，为此将具有相似内容请求偏好的用户划分为同一兴趣社区，如图1所示，具有相同颜色的终端用户处于同一兴趣社区，缓存用户只需缓存所处兴趣社区请求概率最高且自己感兴趣的内容集，当普通用户请求内容时，可通过D2D的方式进行内容分享．

假设用户与用户之间的平均传输时延为T_{d1}，用户到微基站的平均传输时延为T_{d2}，微基站到内容服务器的平均传输时延为T_{d3}．则用户获取内容的平均传输时延T_1为：

$$T_1=\frac{n_1T_{d1}+n_2T_{d2}+(n-n_1-n_2)T_{d3}}{n}\qquad(1)$$

其中，$T_{d1}<T_{d2}<T_{d3}$，n_1表示从缓存用户获取内容的数量，n_2表示从微基站中获取内容的数量，n表示用户获取内容的总数量．

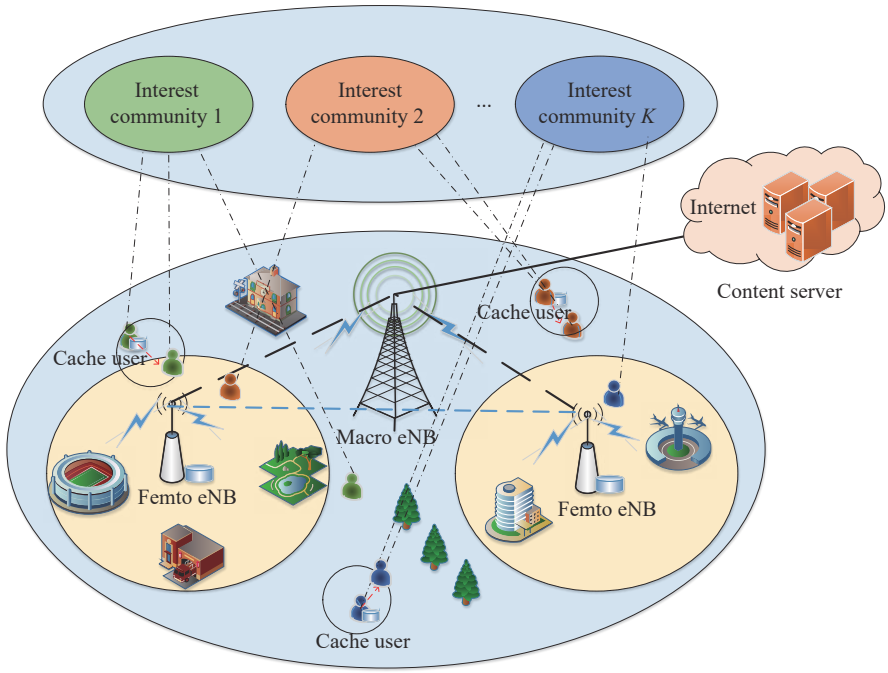

图 1　用户属性感知的边缘缓存模型

Fig.1　User-aware edge-caching model

为减少传输时延, 提高用户体验质量, 从公式(1)可推出, 需要尽可能将用户所需内容缓存, 为此需要在有限的缓存容量下, 提高微基站与缓存用户的缓存命中率. 为提高缓存命中率, 需要进行本地流行内容预测, 以便将用户请求概率最高的内容集缓存, 与此同时, 考虑到 D2D 的传输距离以及用户的自私性, 需要研究缓存用户的选择及其内容的缓存问题以保证用户所需内容的成功缓存与传递. 因此本地流行内容与终端用户缓存策略对提高缓存命中率、减少传输时延、提高用户体验质量至关重要.

2　本地流行内容预测

在传统的无线网络中, 通常是各基站独立地缓存最流行的内容, 但由于用户的兴趣不同, 同一内容对不同用户有着不同的兴趣度, 因此缓存最流行的内容不仅不能满足所有用户的偏好, 还会存在缓存冗余, 浪费缓存资源的现象. 为满足用户偏好, 提高缓存命中率, 减少缓存冗余, 本文从用户的角度出发, 预测出本地流行内容, 并将之协作缓存. 本地流行内容预测分为内容请求概率预测、协作缓存、缓存内容更新.

2.1　内容请求概率预测

本文使用隐语义模型来预测内容请求概率, 基于用户的偏好, 通过隐含特征将用户的兴趣和内容联系起来, 从而获知用户对每类内容的兴趣度, 进而预测出用户对所有内容的请求概率. 不同于传统的人工分类, 隐语义模型是基于用户的行为自动聚类, 因此避免了人工分类的主观性, 提高了预测精度.

在内容服务器处获取一段时间内用户的内容评分矩阵$\boldsymbol{R}$, 其矩阵值r_{uf_i}表示用户u对内容f_i的评分, $r_{uf_i}\in[0,5]$. 通过用户对内容的评分可推出用户对内容的请求概率, 由于用户不可能对内容服务器中所有内容都评过分, 因此矩阵$\boldsymbol{R}$是高度稀疏的. 利用隐语义模型对矩阵进行填充, 可预测出用户对所有内容的评分, 最终预测出用户对所有内容的请求概率. 假设内容服务器中内容的隐含特征有k类, 基于内容在k类隐含特征中所占的权重, 以及用户对k类隐含特征的兴趣度, 可预测出用户对内容的评分$\hat{r}_{uf_i}$:

$$r_{uf_i}\approx\hat{r}_{uf_i}=\boldsymbol{P}_u\boldsymbol{Q}_{f_i}=\sum_{k=1}^{k}p_{uk}q_{kf_i}\tag{2}$$

其中, $\hat{r}_{uf_i}$表示预测的用户 u 对内容f_i的评分, $\boldsymbol{P}_u$是用户–类矢量, 矢量值p_{uk}表示用户 u 对隐含特征k的兴趣度. $\boldsymbol{Q}_{f_i}$是类–内容矢量, 矢量值q_{kf_i}表示内容f_i在隐含特征k中的权重.

为预测用户对所有内容的评分, 需求用户–类矩阵$\boldsymbol{P}=\{\boldsymbol{P}_1^{\mathrm{T}},\boldsymbol{P}_2^{\mathrm{T}},\cdots,\boldsymbol{P}_u^{\mathrm{T}}\}^{\mathrm{T}}$和类–内容矩阵$\boldsymbol{Q}=\{\boldsymbol{Q}_1,\boldsymbol{Q}_2,\cdots,\boldsymbol{Q}_{f_i}\}$中的参数值, 可通过最小化真实值与预测值之间的误差, 即最小化损失函数来求参数值. 损失函数如下:

$$e_{uf_i}^2=\sum_{(u,f_i)\in \boldsymbol{R}}(r_{uf_i}-\hat{r}_{uf_i})^2=\sum_{(u,f_i)\in \boldsymbol{R}}\left(r_{uf_i}-\sum_{k=1}^{k}p_{uk}q_{kf_i}\right)^2+\lambda\|\boldsymbol{P}_u\|^2+\lambda\left\|\boldsymbol{Q}_{f_i}\right\|^2 \tag{3}$$

$\lambda\|\boldsymbol{P}_u\|^2+\lambda\left\|\boldsymbol{Q}_{f_i}\right\|^2$是正则化项，为了防止过拟合. λ是正则化系数，为了控制正则化强弱.

利用随机梯度下降法来最小化损失函数，随机梯度下降法是通过不断判断和选择当前目标下最优的路径，从而在最短路径下达到最优结果，因此，可通过对p_{uk}、q_{kf_i}两个参数求偏导，不断判断和选择损失函数最快的下降方向，p_{uk}和q_{kf_i}对应的偏导数求解如下：

$$\begin{cases}\dfrac{\partial}{\partial p_{uk}}e_{uf_i}^2=-2\left(r_{uf_i}-\sum\limits_{k=1}^{k}p_{uk}q_{kf_i}\right)q_{kf_i}+2\lambda p_{uk}\\ \dfrac{\partial}{\partial q_{kf_i}}e_{uf_i}^2=-2\left(r_{uf_i}-\sum\limits_{k=1}^{k}p_{uk}q_{kf_i}\right)p_{uk}+2\lambda q_{kf_i}\end{cases} \tag{4}$$

不断修改p_{uk}, q_{kf_i}的值，使损失函数的值越来越小，直到收敛为止. 最终使得矩阵$\boldsymbol{P}$和矩阵$\boldsymbol{Q}$的乘积越逼近矩阵$\boldsymbol{R}$. 迭代更新公式如下：

$$\begin{cases}p'_{uk}=p_{uk}+\alpha\left(\left(r_{uf_i}-\sum\limits_{k=1}^{k}p_{uk}q_{kf_i}\right)q_{kf_i}-\lambda p_{uk}\right)\\ q'_{kf_i}=q_{kf_i}+\alpha\left(\left(r_{uf_i}-\sum\limits_{k=1}^{k}p_{uk}q_{kf_i}\right)p_{uk}-\lambda q_{kf_i}\right)\end{cases} \tag{5}$$

式中，α是学习率，决定迭代下降的速率，取值范围为 [0,1].

求得用户对内容的评分矩阵$\hat{\boldsymbol{R}}$后，将分数归一化，即可得用户u对f_i的请求概率$\text{pro}_{u,f_i}=\hat{r}_{uf_i}/5$.

2.2　协作缓存

宏基站较微基站有更大的覆盖区域，可以服务更多用户，但存在信号覆盖的盲区和弱区，微基站通常部署在靠近用户终端，虽然覆盖范围小，但适合小范围精确的覆盖可以弥补宏基站覆盖的不足，通过宏基站的协调将内容缓存在微基站可使用户就近获取缓存资源，减少时延，提高用户体验质量，但由于微基站的缓存容量有限，为充分利用缓存资源，减少缓存冗余，提高缓存内容多样性，增加缓存命中率，需要宏基站小区内的微基站协作缓存. 从 2.1 节可得到宏基站小区内的终端用户对所有内容的请求概率pro_{u,f_i}，将请求概率按降序排列，得到用户i未请求过的前 n 个内容$F_i=\{f_1^i,f_2^i,\cdots,f_n^i\}$，假设宏基站小区内有 m 个终端用户，$s=\{S_1,S_2,\cdots,S_s\}$个微基站，每个微基站的缓存容量相同且为c_{SBS}，每个微基站小区内有m_{S_i}个终端用户，且满足$m=m_{S_1}+m_{S_2}+\cdots+m_{S_s}$. 对微基站$S_i$小区内所有用户未请求过的前 n 个内容求交集$F_1\cap F_2\cap\cdots\cap F_{m_{S_i}}=\{b_{f_1^{S_i}},b_{f_2^{S_i}},\cdots,b_{f_{n_i}^{S_i}}\}$，其中$b_{f_1^{S_i}}$表示内容$f_1$在微基站$S_i$中将请求的次数，$n_i$表示微基站覆盖范围内用户请求的内容总数，在微基站缓存容量的限制下$d\cdot\sum\limits_{i=1}^{s}\sum\limits_{j=1}^{a}f_j^{S_i}\leqslant s\cdot c_{\text{SBS}}$，其中$a$表示在每个微基站中能缓存的最大内容数量，得到微基站小区内请求数量最多的内容集$F_{S_i}=\{f_1^{S_i},f_2^{S_i},\cdots,f_a^{S_i}\}$，满足$b_{f_1^{S_i}}>b_{f_2^{S_i}}>\cdots>b_{f_{n_i}^{S_i}}$，并将之缓存，在宏基站中储存了所管理微基站的缓存列表，宏基站协调微基站进行内容缓存，使之满足每个微基站内缓存的内容无重复$F_{S_1}\cap F_{S_2}\cap\cdots\cap F_{S_s}=\emptyset$，使得缓存资源得到充分利用，并提高了缓存命中率. 综合所有微基站缓存的内容即为本地流行内容$F_{\text{local}}=\{F_{S_1},F_{S_2},\cdots,F_{S_s}\}$.

2.3　缓存内容更新

考虑到用户的偏好可能随时间而不断变化，以及预测结果不完善造成某些请求概率较高的内容未被缓存的情况，因此需要及时更新缓存的本地流行内容.

内容的缓存可分为主动缓存和被动缓存. 主动缓存是指通过预测内容的流行度，提前将流行的内容缓存，被动缓存是指根据内容的历史访问情况决定该内容是否被缓存. 预测的内容流行度与实际内容流行度误差较小时，主动缓存可大大提高缓存效益，但若误差较大，缓存性能将不如被动缓存，被动缓存的缓存效益稳定，但存在较大的提升空间. 为提高缓存性能，结合主动缓存与被动缓存的优点，本文将微基站分为两个缓存区：c_1 缓存区和 c_2 缓存区，两缓存区的总存储容量为c_{SBS}，两缓存区的大小可自适应调整. 在 c_1 区进行主动缓存，其内容缓存时间不超过一个有限时间T_{cache}，在 c_2 区进行被动缓存，其内容缓存时间超过一个有限时间T_{cache}. 内容缓存到微基站后，微基站实时记录每个用户的内容请求，并在一个时间段$t(t<T_{\text{cache}})$内计算内容的实际流行度，若在一个时间段 t 内，某个未缓存的内容f_i实际流行度大于缓存区的最低实际流行度，则需要及时更新本地流行内容，将实际流行度最低的内容替换为内容f_i，每隔一个有限的时间范围T_{cache}，删除实际流行度低的内容，并将 c_1 区和 c_2 区实际流行度最高的内容集移动到 c_2 区，然后再重新预测用户的内容请

求概率, 更新本地流行内容, 并将之缓存在 c_1 区. 一个时间段 t 内, 内容 f_i 的实际流行度(即内容被请求的概率) $\text{pro}_{t,f_i}^{\text{real}}$ 为:

$$\text{pro}_{t,f_i}^{\text{real}} = (\sum_{j=1}^{t} T_j^{f_i})/t \tag{6}$$

其中, $\sum_{j=1}^{t} T_j^{f_i}$ 表示第 t 个时间段内内容 f_i 被请求的次数.

3　终端用户缓存策略

随着缓存以及 D2D 技术的发展, 终端用户可以不依赖于基站就能直接进行内容分享, 为进一步减少传输延迟, 本文在这部分研究终端用户的缓存策略, 分为兴趣社区划分、缓存用户确定、内容缓存.

3.1　兴趣社区划分

考虑到用户的自私性可能不愿意缓存其不感兴趣的内容, 因此需要将兴趣相似的用户划分为一个兴趣社区, 使缓存用户只需为同一兴趣社区的用户缓存内容, 这可以提高用户的缓存意愿. 从 2.1 节可得到每个用户对隐含特征的兴趣度 $\boldsymbol{P}_u = \{p_{u1}, p_{u2}, \cdots, p_{uk}\}$, 使用欧式距离来衡量用户间的兴趣相似度, 距离与相似度成反比, 距离越近相似度越高.

$$\text{dist}(u_i, u_j) = \sqrt{\sum_{k=1}^{k} (p_{u_i k} - p_{u_j k})^2} \tag{7}$$

$$S_{u_i,u_j}^{\text{interest}} = 1/(1 + \text{dist}(u_i, u_j)) \tag{8}$$

其中, $\text{dist}(u_i, u_j)$ 表示用户 u_i,u_j 之间的距离, $S_{u_i,u_j}^{\text{interest}}$ 表示用户 u_i,u_j 之间的兴趣相似度.

由于 k-means++聚类算法具有对数据进行分析并自动聚类的特点, 并且有效缓解了 k-means 聚类算法对初始中心选取不当而陷入局部最优的情形, 因此, 本文选用 k-means++算法对用户进行聚类, 进而形成兴趣社区. 首先, 随机选择一个用户作为第一个聚类中心 μ_1, 然后计算所有用户到 μ_1 的距离, 选择距离最大的用户作为第二个聚类中心 $\mu_2 = \arg\max_{j\in\{2,\cdots,m\}} \text{dist}(u_j, u_1)$, 以此类推, 选出 K 个聚类中心. 然后, 计算每个用户 u_j 到 K 个聚类中心 μ_K 的距离, 用户 u_j 将与自己最近的聚类中心 μ_K 聚类进而形成兴趣社区 $C_k = \{j:\ K = \arg\min_{i\in\{1,2,\cdots,K\}} \text{dist}(u_i, u_j)\}$, 每个用户找到自己的兴趣社区后, 再重新计算新的聚类中心 $\mu_K = \dfrac{1}{|C_K|} \sum_{u\in C_K} \boldsymbol{P}_u$, 重复, 直到聚类中心不再有变化或达到最大迭代次数.

3.2　缓存用户确定

在每个兴趣社区 C_K 中, 基于用户的缓存意愿以及缓存能力来确定缓存用户.

在移动社交网络中, 用户的缓存意愿与用户间的社交关系密切相关, 通过考虑缓存用户与普通用户间的信任度与社交相似度, 来确定其缓存意愿. 信任度是用来评估缓存用户是否愿意将内容分享给普通用户的信任程度. 在一段时间内, 假设用户 u_b 从用户 u_a 中成功获取内容的次数为 $\text{Num}_{u_\text{b},u_\text{a}}$, 从所有用户中成功获取内容的次数为 $\text{Num}_\text{b} = \sum_{j\in C_K} \text{Num}_{u_\text{b},u_j}$, 则用户 u_b 对 u_a 的信任度为 $\text{Trust}_{u_\text{b},u_\text{a}} = \text{Num}_{u_\text{b},u_\text{a}}/\text{Num}_\text{b}$, 综合兴趣社区 C_K 中所有用户对 u_a 的信任度, 即可得到用户 u_a 的信任值 $\text{Trust}_{u_\text{a}} = \sum_{i\in C_K} \text{Trust}_{u_i,u_\text{a}}$, 将用户的信任值范围限定到 0 到 1 之间, 即可得到用户 u_a 的信任概率 $\text{pro}_{u_a}^{\text{trust}}$:

$$\text{pro}_{u_a}^{\text{trust}} = \frac{1+\text{e}^{-1}}{1-\text{e}^{-1}} \cdot \left(\frac{2}{1+\text{Trust}_{u_\text{a}}} - 1\right) \tag{9}$$

社交相似度是指用户社会背景(年龄、性别、职业等)的相似度, 用户间的社交相似度越高, 则用户为其缓存内容的意愿越强. 设 $\vec{I}_\text{a} = \{I_\text{a1}, I_\text{a2}, \cdots I_\text{aL}\}$ 为用户 u_a 的社交特征向量, 其中 L 表示用户社交特征的数量, 使用余弦相似度来衡量用户间的社交相似度 $S_{u_\text{a},u_\text{b}}^{\text{social}}$:

$$S_{u_\text{a},u_\text{b}}^{\text{social}} = \frac{\vec{I}_\text{a} \cdot \vec{I}_\text{b}}{\left\|\vec{I}_\text{a}\right\| \times \left\|\vec{I}_\text{b}\right\|} = \frac{\sum_{k=1}^{L} (I_{\text{a}k} \times I_{\text{b}k})}{\sqrt{\sum_{k=1}^{L} (I_{\text{a}k})^2} \times \sqrt{\sum_{k=1}^{L} (I_{\text{b}k})^2}} \tag{10}$$

综合兴趣社区 C_K 中所有用户与 u_a 的社交相似度, 即可得到用户 u_a 的社交相似度 $S_{u_\text{a}}^{\text{social}} = \sum_{u_\text{b}\in C_K} S_{u_\text{a},u_\text{b}}^{\text{social}}$, 将社交相似度范围限定到 0 到 1 之间, 即可得到用户 u_a 的社交概率 $\text{pro}_{u_\text{a}}^{\text{social}}$:

$$\text{pro}_{u_\text{a}}^{\text{social}} = \frac{2}{\pi} \arctan(S_{u_\text{a}}^{\text{social}}) \tag{11}$$

在移动社交网络中, 由于用户的移动性, 可能使两个用户不在 D2D 通信范围内, 最终导致无法进行内容分享, 因此, 除了考虑用户的缓存意愿还需考虑用户的缓存能力, 用户的缓存能力可用用户的交互相似度来衡量. 交互相似度是由相遇平均间隔、相遇平均持续时间共同决定.

用户在一段时间内的相遇间隔越短, 则分享内容的概率越大. 若在时间段 T 内用户 u_a 和用户 u_b 第 i 次交互过程的起始时刻和终止时刻为 T_i^1 和 T_i^2,

交互次数为n_{inter}, 两个用户在时间段 T 内第 i 次和第 i+1 次的交互过程的时间间隔Δt_i为$T_{i+1}^1 - T_i^2$, 则在时间段 T 内, 两个用户的相遇平均间隔$T_{u_\text{a},u_\text{b}}^{\text{inter_avg}}$为:

$$T_{u_\text{a},u_\text{b}}^{\text{inter\_avg}} = \text{E}(\Delta t_i) = \sum_{i=1}^{n}(T_{i+1}^1 - T_i^2)/n_{\text{inter}} \tag{12}$$

用户在一段时间内的相遇持续时间越长, 内容分享成功的概率越高, 若两个用户在时间段 T 内第 i 次交互的持续时间T_{dur}^i为$T_i^2 - T_i^1$, 距离当前时刻近的交互过程更能准确反应用户间的交互关系, 使用排序加权权重法[11]来得到时间段 T 内两个用户间的相遇平均持续时间$T_{u_\text{a},u_\text{b}}^{\text{dur_avg}}$为:

$$T_{u_\text{a},u_\text{b}}^{\text{dur\_avg}} = \sum_{i=1}^{n_{\text{inter}}} \xi_i \cdot T_{\text{dur}}^i \tag{13}$$

其中, ξ_i为第 i 次交互的相对权重, $\xi_i = 2i/n_{\text{inter}}(n_{\text{inter}}+1)$.

用户间的交互相似度随着相遇平均间隔$T_{u_\text{a},u_\text{b}}^{\text{inter_avg}}$值的减小及相遇平均持续时间$T_{u_\text{a},u_\text{b}}^{\text{dur_avg}}$的增加而增加,并逐渐趋于稳定. 交互相似度的取值范围为 [0,1], 0 表示交互相似度最低, 1 表示交互相似度最高, 进而计算出交互相似度$S_{u_\text{a},u_\text{b}}^{\text{inter}}$为:

$$S_{u_\text{a},u_\text{b}}^{\text{inter}} = \frac{1}{1+\exp\left(-\dfrac{T_{u_\text{a},u_\text{b}}^{\text{dur\_avg}}}{T_{u_\text{a},u_\text{b}}^{\text{inter\_avg}}}\right)} \tag{14}$$

综合兴趣社区 C_K 中所有用户与u_a的交互相似度, 即可得到用户u_a的交互概率$\text{pro}_{u_\text{a}}^{\text{inter}}$:

$$\text{pro}_{u_\text{a}}^{\text{inter}} = \sum_{b\in C_K} S_{u_\text{a},u_\text{b}}^{\text{inter}}/m_{C_K} \tag{15}$$

其中, m_{C_K}表示兴趣社区 C_K 中用户的总数.

根据用户的缓存意愿和缓存能力, 最终得到用户作为缓存用户的概率pro_{u_a}:

$$\text{pro}_{u_\text{a}} = \text{pro}_{u_\text{a}}^{\text{trust}} \cdot \text{pro}_{u_\text{a}}^{\text{social}} \cdot \text{pro}_{u_\text{a}}^{\text{inter}} \tag{16}$$

最后, 将兴趣社区 C_K 中, 用户的缓存概率pro_{u_a}按降序排列, 选择缓存概率最高的前m_{UE}个终端用户作为缓存用户.

3.3　内容缓存

考虑到用户的自私性以及有限的缓存资源, 在确定缓存用户后, 为提高缓存命中率、减小传输时延, 需要为其分配合适的内容来缓存. 从 2.2 节可得终端用户未请求内容中请求概率最高的内容集$F_i = \{f_1^i, f_2^i, \cdots, f_n^i\}$, 综合兴趣社区$C_K$中所有用户的内容请求概率, 按请求次数降序排列, 最终得到兴趣社区C_K中请求概率最高的内容集$F_{C_K} = \{f_1^{C_K}, f_2^{C_K}, \cdots, f_n^{C_K}\}$, 将缓存用户请求概率最高的内容集$F_i$以及兴趣社区$C_K$中请求概率最高的内容集$F_{C_K}$求交集, 将交集内容中请求次数最高的内容集$F_i^{\text{cache}}$发送给缓存用户使之缓存. 缓存用户$u_i$缓存的内容$F_i^{\text{cache}}$为:

$$\begin{aligned} &F_i^{\text{cache}} = F_i \cap F_{C_K} = \{f_1^{\text{cache}}, f_2^{\text{cache}}, \cdots, f_j^{\text{cache}}\} \\ &\text{s.t.} \begin{cases} b_{f_1^{\text{cache}}} > b_{f_2^{\text{cache}}} > \cdots > b_{f_j^{\text{cache}}} \\ j \cdot d \leqslant c_{\text{UE}} \end{cases} \end{aligned} \tag{17}$$

其中, $b_{f_1^{\text{cache}}}$表示内容f_j^{cache}将被请求的次数, j表示内容数, c_{UE}表示终端用户的存储容量, 限制条件满足, 缓存用户缓存的内容大小小于其缓存容量.

4　数值结果分析

本节在 MATLAB 平台下, 对所提出的用户属性感知的边缘缓存机制(UAEC)的性能进行分析, 仿真参数设置如下, 考虑一个宏基站小区内的边缘缓存场景, 由 6 个微基站和 78 个终端用户组成, 采用 INFOCOM2006 会议的实测数据[12]来分析终端用户的移动特性, 采用 GroupLens 小组提供的 MovieLens 的数据集[13]来分析终端用户的偏好以及社会属性, 隐语义模型使用的正则化系数λ为 0.11, 学习率a为 0.1, 隐含特征k为 20, 总的内容数量为 1682, 并假设内容服务器到终端用户的平均传输时延为 98 ms, 微基站到终端用户的平均传输时延为 10 ms, 终端用户到终端用户的平均传输时延为 3 ms, 微基站的覆盖半径为 100 m, D2D 的通信距离为 30 m, 微基站和缓存用户的缓存容量的比例为$\{c_{\text{SBS}}, c_{\text{UE}}\} = \{30 \cdot c_{\text{UE}}, c_{\text{UE}}\}$.

4.1　性能分析

为分析本地流行内容预测和缓存用户的选择对所提机制的性能影响, 将所提机制(UAEC)与下面两种情况在缓存命中率方面进行对比, 缓存命中率是指缓存的内容请求数占总的内容请求数的比例.

1)没有本地流行内容预测(Without local popular prediction, WLPP): 不根据用户的偏好对本地流行内容进行预测, 并假设内容流行度服从 Zipf 分布$\text{pro}_{f_j} = (1/j)^\beta / \sum_{k=1}^{N}(1/k)^\beta$, 其中 β表示内容流行度偏度[14].

2)没有缓存用户的选择(Without cache users selection, WCUS): 不考虑终端用户的缓存意愿与缓存能力, 而是以相同的概率来选择缓存用户.

由图 2 可知, 随着缓存容量的增加, 所提机制

与其它两种情况的缓存命中率均呈现上升趋势, 这是因为随着缓存容量的增加, 在微基站和缓存用户中缓存的内容数增加, 使更多用户的需求得到满足. 而本文所提机制相比其他两种情况有更高的缓存命中率, 这是因为由于用户的自私性, 可能不愿意占用自己的缓存资源来缓存不感兴趣的内容, 若让缓存用户缓存其不感兴趣的内容, 可能造成内容丢弃, 最终导致较低的缓存命中率, 与此同时, 若缓存用户与请求者不在 D2D 通信范围内, 将不能进行内容分享, 最终也导致较低的缓存命中率, 而所提机制基于用户兴趣, 构建兴趣社区, 在每个兴趣社区中基于用户的缓存意愿以及缓存能力, 对缓存用户进行选择进而提高缓存命中率. 对于内容流行度, 若考虑所有用户的内容流行度服从同一分布, 也将造成较低的缓存命中率, 这是因为内容流行度对不同用户是不一样的, 为此, 所提机制基于用户的偏好, 来预测用户的内容流行度, 从而提高了缓存命中率.

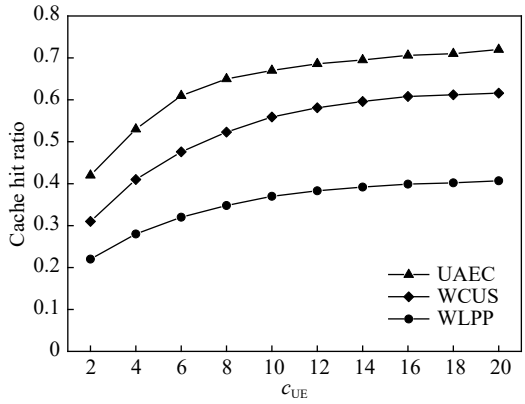

图 2　不同缓存容量下的缓存命中率

Fig.2　Cache hit ratio versus cache capacity

图 3 展示了缓存用户数对缓存命中率的影响, 从图中可以看出, 随着缓存用户数的增加, 缓存命中率逐渐增加, 这是因为内容总数不变的情况下, 更多的缓存用户意味着有更多的缓存容量, 因此可以为内容请求者缓存更多其感兴趣的内容. 其中所提机制 UAEC 和 WLPP 机制比 WCUS 机制变化的更快, 是因为这两种机制对缓存用户进行了选择, 相比于 WLPP 机制和 WCUS 机制, 本文所提机制实现更高的缓存性能, 当缓存用户数为 20 时, 比没有选择缓存用户和没有预测本地流行内容的情况, 在缓存命中率方面分别提高了 13% 和 30%.

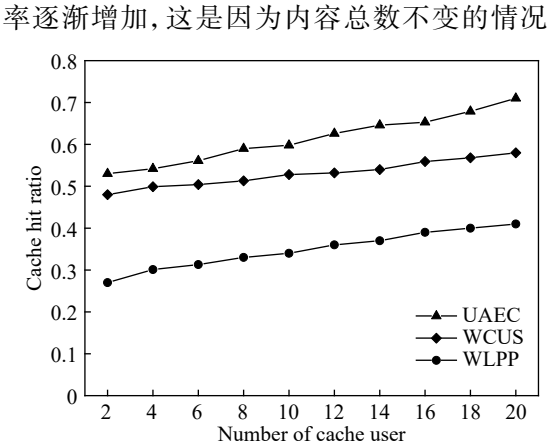

图 3　不同缓存用户数下的缓存命中率

Fig.3　Cache hit ratio versus number of cache users

4.2　与其他机制的比较

为证明本文所提机制的性能, 将其与现存在的两种机制进行比较:

1) 最流行内容缓存[15](Most popular caching, MPC): 在每个微基站和缓存用户中都缓存最流行的内容, 且随机选择缓存用户.

2) 随机缓存[16](Random caching, RC): 微基站和缓存用户以相同的概率缓存所有内容, 且随机选择缓存用户.

由图 4 可知, 随着微基站和缓存用户缓存容量的增加, 所提机制 UAEC 与其它两种机制 MPC 和 RC 的缓存命中率都逐渐增加, 相比于其他两种机制, 所提机制实现更好的性能, 这是因为 MPC 机制, 只考虑缓存最流行的内容, 忽略了用户的偏好, 不能满足部分用户的需求, 同时, 在微基站和缓存用户都缓存最流行内容, 会造成缓存冗余, 减少缓存内容多样性的现象, 最终使缓存命中率较低, 对于 RC 机制, 没有考虑用户的偏好, 随机缓存内容, 因此有较低的缓存命中率, 当缓存容量为 20 时, 所提机制的缓存命中率比 MPC 机制和 RC 机制分别提高了 19.5% 和 47%.

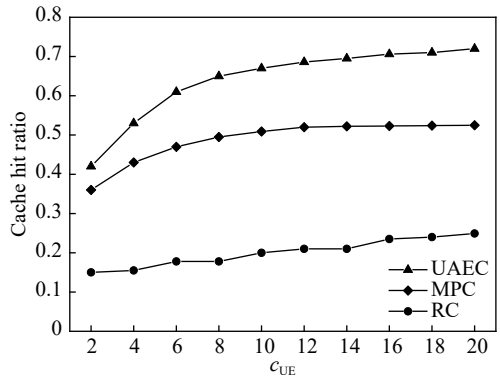

图 4　不同机制中缓存命中率随着缓存容量的变化

Fig.4　Cache hit ratio versus cache capacity in different mechanisms

图 5 展示了不同机制下缓存用户数对缓存命中率的影响, 由图可知, 随着缓存用户数的增加, 三种机制的缓存命中率都逐渐增加, 这是因为随着缓存用户数的增加使总的缓存容量增加, 缓存

用户可以缓存更多的内容来满足请求者的需求. 其中所提机制比其他两种机制的缓存命中率更高, 是因为所提机制基于用户的缓存意愿和缓存能力来选择合适的用户作为缓存用户, 提高了请求者获取感兴趣内容的概率, 因此有更高的缓存命中率, 当缓存用户数为 20 时, 所提机制的缓存命中率比 MPC 机制和 RC 机制分别提高了 29.5% 和 40%.

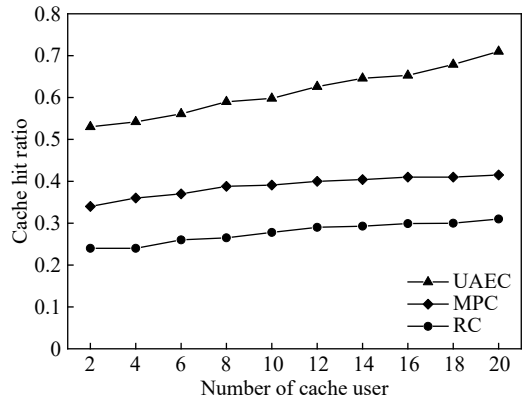

图 5　不同机制中缓存命中率随着缓存用户数变化

Fig.5　Cache hit ratio versus number of cache users in different mechanisms

图 6 展示了不同机制下缓存容量对平均传输时延的影响, 三种机制的平均传输时延随着缓存容量的增加而减小, 其原因在于随着缓存容量增加使得缓存命中率增加, 用户可以从微基站和缓存用户中更多的获取所需内容, 减少了从用户到微基站及微基站到内容服务器间的传输时延, 因此降低了平均传输时延. 由图可知, 本文所提机制比 RC 机制和 MPC 机制整体上有更低的平均传输延迟, 是因为相比于其它两种机制本文所提机制针对用户的偏好, 来进行内容缓存, 同时考虑了用户的缓存意愿和缓存能力来选择缓存用户, 增大用户从缓存用户获取内容的概率, 因此有较低的平均传输时延.

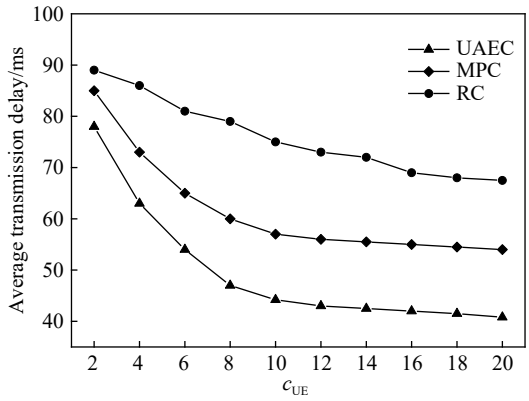

图 6　不同缓存容量下的平均传输时延比较

Fig.6　Average transmission delay versus cache capacity

5　结论

针对数据流量爆发式增长所引发的网络拥塞、用户体验质量恶化等问题, 提出一种用户属性感知的边缘缓存机制. 基于用户偏好预测出本地流行内容, 通过微基站将之协作缓存, 并根据用户偏好变化, 将之实时更新. 为进一步减少传输时延, 基于用户偏好构建兴趣社区, 在每个兴趣社区中基于用户的缓存意愿和缓存能力选择合适的缓存用户来缓存目标内容. 结果表明, 所提机制能够有效的提高缓存命中率, 降低平均传输时延, 提高用户体验质量. 本文虽然考虑的是单个宏基站覆盖范围内的缓存方案, 但对多个宏基站情况下, 所提机制也同样适用, 并且通过考虑宏基站间的协作缓存, 可进一步减少传输时延, 提高用户体验质量.

参　考　文　献

[1] Cisco. Cisco annual internet report (2018 –2023) white paper [R/OL]. *Cisco* (2020-03-09)[2020-05-15]. https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html
(思科. 思科年度互联网报告(2018-2023)白皮书[R/OL]. 思科(2020-03-09)[2020-05-15]. https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html)

[2] Wu D P, Zhang F, Wang H G, et al. Security-oriented opportunistic data forwarding in mobile social networks. *Future Generation Comput Syst*, 2018, 87: 803

[3] Cai J L Z, Yan M Y, Li Y S. Using crowdsourced data in location-based social networks to explore influence maximization // *IEEE INFOCOM* 2016-*The 35th Annual IEEE International Conference on Computer Communications*. San Francisco, 2016: 1

[4] Gregori M, Gómez-Vilardebó J, Matamoros J, et al. Wireless content caching for small cell and D2D networks. *IEEE J Sel Areas Commun*, 2016, 34(5): 1222

[5] Jiang X W, Zhang T K, Zeng Z M. Content clustering and popularity prediction based caching strategy in content centric networking // 2017 *IEEE* 85*th Vehicular Technology Conference* (*VTC Spring*). Sydney, 2017: 1

[6] Zhang Y R, Pan E T, Song L Y, et al. Social network aware device-to-device communication in wireless networks. *IEEE Trans Wireless Commun*, 2015, 14(1): 177

[7] Chen M Z, Saad W, Yin C C, et al. Echo state networks for proactive caching in cloud-based radio access networks with mobile users. *IEEE Trans Wireless Commun*, 2017, 16(6): 3520

[8] Cheng Y Q, Wu M Q, Zhao M, et al. Socially-aware NodeRank-based caching strategy for Content-Centric Networking // 2016 *International Symposium on Wireless Communication Systems* (*ISWCS*). Poznan, 2016: 297

[9] Zirak M, Yaghmaee M H, Tabbakh S R K. A distributed cache points selection scheme for reliable transport protocols with intermediate caching in Wireless Sensor Networks // 16*th International Conference on Advanced Communication Technology*. Pyeongchang, 2014: 705

[10] Al Ridhawi I, Al Ridhawi Y. A cache-node selection mechanism for data replication and service composition within cloud-based systems // 2017 *Ninth International Conference on Ubiquitous and Future Networks* (*ICUFN*). Milan, 2017: 726

[11] Cui L Z, Dong L Y, Fu X H, et al. A video recommendation algorithm based on the combination of video content and social network. *Concurrency Comput*: *Pract. Exper*, 2017, 29(14): e3900

[12] Qiu L, Cao G H. Cache increases the capacity of wireless networks // *IEEE INFOCOM* 2016-*The* 35*th Annual IEEE International Conference on Computer Communications*. San Francisco, 2016: 1

[13] Harper F M, Konstan J A. The movielens datasets: History and context. *ACM Trans Interactive Intell Syst*, 2016, 5(4): 19

[14] Bastug E, Bennis M, Debbah M. Living on the edge: The role of proactive caching in 5G wireless networks. *IEEE Commun Mag*, 2014, 52(8): 82

[15] Ahlehagh H, Dey S. Video-aware scheduling and caching in the radio access network. *IEEE/ACM Trans Networking*, 2014, 22(5): 1444

[16] Blaszczyszyn B, Giovanidis A. Optimal geographic caching in cellular networks // 2015 *IEEE International Conference on Communications* (*ICC*). London, 2015: 3358