applied sciences

MDPI

Article

User-In-The-Loop Content Delivery in Cellular Communication Networks with Heterogeneous User Behaviors

Weicheng Zhang [1], Yaodong Li [1], Hai Lu [1,*], Xuemin Hong [1,2] and Jianghong Shi [1,2]

[1] School of Information Science and Technology, Xiamen University, Xiamen 361005, China; 19720152203547@stu.xmu.edu.cn (W.Z.); liyaodongwork@163.com (Y.L.); xuemin.hong@xmu.edu.cn (X.H.); Shijh@xmu.edu.cn (J.S.)

[2] Key Lab of Underwater Acoustic Communication and Marine Information, Ministry of Education, Xiamen University, Xiamen 361005, China

* Correspondence: luhai@xmu.edu.cn; Tel.: +86-592-258-0150

Received: 25 March 2018; Accepted: 23 April 2018; Published: 2 May 2018

Abstract: User-in-the-loop (UiL) content delivery is a recently proposed scheme for personalized content retrieval over mobile communication networks. It is a promising scheme that can better manage the overall user quality-of-experience (QoE) throughout the entire content retrieval process. The performance of the scheme, however, has only been investigated in a simplified system model that assumed synchronized user behavior, inflexible delay constraint, and identical quality-of-service (QoS) requirement among users. This paper studies the performance of UiL content delivery scheme in a generalized and realistic system model with asynchronous user behavior, flexible outage delay constraint, and customized user QoS. Heuristic algorithms and theoretical bounds are investigated for the UiL content delivery problem. The proposed system is shown to be effective in managing the user QoE in generalized practical scenarios.

Keywords: content recommendation; content delivery; mobile communication; QoS; QoE

1. Introduction

Measurements revealed that content retrieval applications (e.g., website browsing, video streaming, and file download) have contributed to the majority of traffic running on the mobile Internet. To help users obtain the most relevant and timely information, personalized content retrieval services have recently emerged as a novel type of content retrieval service. Personalized content retrieval relies on the historical user-behavior data to drive recommendation algorithms [1,2], which can effectively process a vast content pool and select a small subset of content that would best appeal to the targeted user's interest. As an indispensable tool in the era of information explosion, personalized content retrieval technology has attracted wide research interests worldwide.

The service of personalized content retrieval is provided in two consecutive phases: content recommendation phase and content delivery phase. The first phase is responsible for processing a vast amount of user-relevant contextual data [3,4], such as historical online behavior, social relationships, and mobility patterns, to establish user profiles. The information of user profiles and content properties are then jointly processed to rate the user's potential interest on a piece of content. Contents with the highest rating can then be recommended to the user. The latter phase of content delivery is responsible for end-to-end delivery of the requested content via the communication network with certain quality-of-service (QoS) requirement. In the personalized content retrieval service, user's quality-of-experience (QoE) is related to both phases. First, at the semantic level, the recommended content should be appealing to users; second, at the perception level, the content

download/viewing process should be smooth, which translates to high communications throughput and low delay.

Nowadays, content recommendation and content delivery has evolved into two different industries. Players of the content recommendation industry are called over-the-top players (OTTs) [5]. These players own the (copyright of) contents and track the users' content preferences using big-data technologies. On the other hand, players of the content delivery industry are called content delivery networks (CDNs). These players build general-purpose communication infrastructures for a variety of applications (including but not limited to content delivery applications). Although such a separation is intuitive and convenient, recent studies have shown that once the tasks of content recommendation and content delivery can be jointly optimized, i.e., if OTTs and CDNs are allowed to cooperate or jointly designed, the overall performance of content retrieval can be improved significantly [6,7]. The rationale of joint OTT and CDN design is to consider the human/user factor in the content delivery, thus is also called user-in-the-loop (UiL) content delivery. This concept has inspired a wealth of research in recent years.

Most studies about UiL content delivery in the literature focused on a design paradigm called proactive caching [8–10]. Proactive caching selects a set of recommended content, pushes the content through the network when there is available capacity, and caches the contents at the user device before a user request actually happens. The proactive caching scheme has been systematically investigated in the literature, covering many issues such as multicast streaming [11,12], energy and bandwidth efficiency [13,14], heterogeneous overlay networks [12], dynamic traffic pricing [15,16] and D2D communication networks [17–19]. Although the proactive caching technology does help in network load balancing and service delay reduction, it has some inherent shortcomings. First, the user devices should have a large amount of storage space, resulting in high costs and energy consumption. Second, if the pushed content is not viewed by the users eventually, the network resources used for delivering the content is wasted. Third, it is difficult to measure how much traffic volume is actually used by the user, hence traffic pricing could be problematic.

As an improvement to proactive caching, we recently proposed a new protocol of UiL content delivery in [20]. Figure 1 illustrates the four steps of the proposed protocol. (1) The base station (BS) collects relevant physical layer information and send it to the central server to indicate the capacity of each radio link; (2) Taking the physical layer information as a key reference, the central server generates a personalized list of content recommendation for each user; (3) Users browse the recommended list and request interesting content from the list; (4) Content is delivered from the central server to the user. The novelty of our protocol lies in that content recommendation is used as a traffic shaping technique to avoid traffic congestion and excessive delay. Unlike the proactive caching protocols, which is a "push-type" protocol, our protocol is a "pull-type" protocol in a sense that it only delivers a file after the actual user request occurs. As a result, all traffic in the proposed system is on-demand traffic requested by the user, so that the problem of invalid traffic no longer exists. Moreover, in comparison with the conventional hypertext transport protocol (http), which is also a pull-type protocol, our protocol is more advanced in that it exploits the semantic level information to jointly optimize the overall user experience and reduces the risk of network congestion.

Our previous work in [20] has three limitations. First, we adopted a simplified user behavior model, which assumed that users' requests for content occur at the same time. This synchronous user behavior model corresponds to the worst-case scenario when the content retrieval system always generates the highest traffic demand. In practice, however, the worst-case scenario will only occur infrequently; therefore, using the worst-case model will result in unnecessarily conservative system performance in practice. Second, the delay performance is not thoroughly investigated. Our previous work adopted a simplified assumption that all content requests should be delivered within a time slot. This means enforcing a hard and identical threshold on the content delivery delay. In practice, however, the content delivery service is not a time-critical service and is allowed to have elastic delays in most applications' scenarios. It is therefore desirable to study the system performance using a different

model of outage delays. Third, all users are treated equally in terms of QoS requirements, which means that the delay requirements are identical for all users. In practice, it is desirable to have the flexibly to customize the QoS for different users.

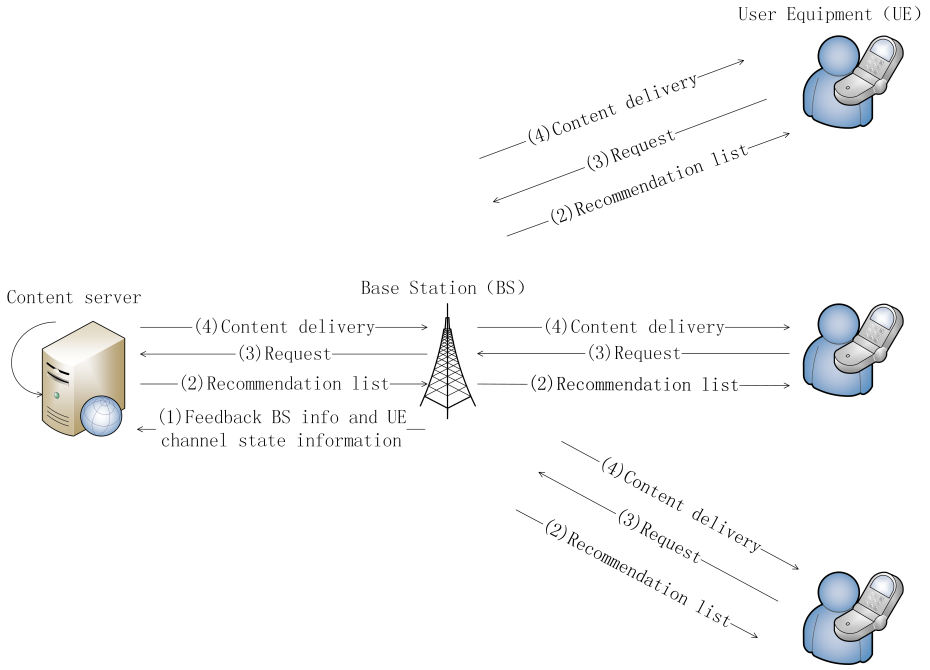

Figure 1. Interactive content retrieval protocol for personalized multi-user content retrieval.

This paper contributes in improving our previous work in [20] by addressing the above-mentioned three limitations. First, the synchronous user behavior model is extended to an asynchronous model, which better captures the user behaviors in reality. Second, we use outage delay to bound the delay performance of the content delivery system. The outage delay constraint is a generalization from the fixed delay constraint. Third, our algorithm in this paper can support customized user QoS by allowing different delay constraints for different users. Hence, it allows more flexibility to manage the overall QoE in practice. In summary, our work generalizes the assumptions of homogeneous user behavior, delay constraint, and user QoS requirement to the heterogeneous cases.

The rest of this paper is organized as follows. Section 2 introduces the system model and formulates the problem of UiL content delivery. Derivations of several theoretical performance bounds are presented in Section 3. A heuristic algorithm to solve the problem is proposed in Section 4, followed by simulation results in Section 5. Finally, conclusions are drawn in Section 6.

2. System Model

2.1. System Architecture

The architecture of a UiL content delivery system is illustrated in Figure 1. The system includes three types of entities: central server, BSs, and users. Each user is associated with one BS, while each BS is connected to the central server, which stores a total of F content files. It is assumed that each BS, once allocated with a fixed amount of radio resources, operates independently in performing the task of content delivery, so that our study can focus on a single BS. The number of users in the considered BS is denoted by U.

Using existing recommendation algorithms [21–24], the central server tracks the interest profile of all users in the BS and outputs a $U \times F$ user interest matrix $\mathbf{R}$. The elements of $\mathbf{R}$ is denoted by r_{uf}, which indicates the interest of the u-th user in the f-th file. Here, $u(u = 1, 2, \ldots, U)$ and $f(f = 1, 2, \ldots, F)$ are the indexes for users and content files, respectively. We use L_f to denote the size

of the f-th content file. The file size varies across different content. For simplicity, it is assumed that the user interest profile is independent from the file size. The UiL content delivery system aims to maximize the sum interests of all contents recommended to all users in the cell, under the condition that the recommended contents can be delivered to the user in time with a predefined probability.

2.2. Traffic Demand Model

The asynchronous user behavior and the resulted traffic demand model are illustrated in Figure 2. The time is divided into frames of duration T. Each time frame is further divided into n time slots. The list of content recommendation is updated at the beginning of each time frame. Within the time frame, we assume that users will browse the recommendation list and request content from the list in an asynchronous manner. It is assumed that the behavior for user u to request a content at each time slot follows a Bernoulli process with parameter λ_u, which means the probability for user u to place a content request at the time slot t is

$$P_{u,t}(x) = \begin{cases} \lambda_u, & x = 1, \\ 1 - \lambda_u, & x = 0. \end{cases} \tag{1}$$

This model allows the user requests to be asynchronous in time and is a generalization of the model in [20]. When $\lambda_u = 0$, it means that user u is inactive and does not request any content at the time period. When $\lambda_u = 1$, it means that the user always requests one and only one piece of content at each time slot. In this case, our model reduces to the extreme case in [20]. The arrival rate λ_u is assumed to be stable during a time frame but can vary across different time frames and different users.

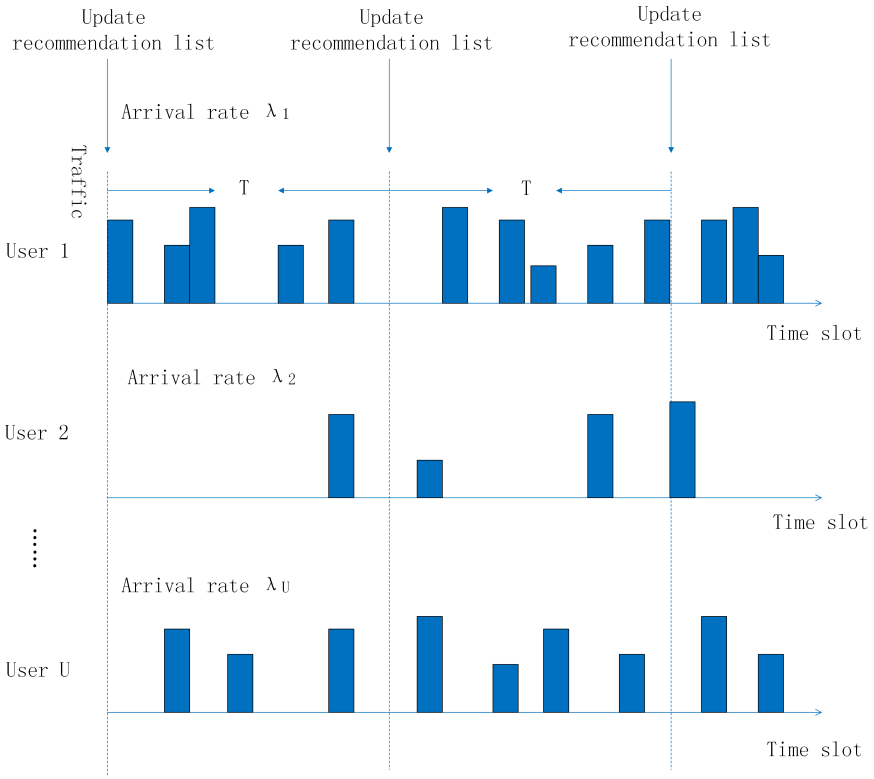

Figure 2. Illustration of the traffic demand model.

Two general assumptions are further made about the traffic demand model. First, we assume that the size of a common content file (which can be a video) is much larger than the size of a recommendation list (which includes mainly text and small image), so that the transmission time of

the recommendation list is negligible. Second, we assume that the large scale (i.e., slow fading) channel gains remain unchanged during a cycle of content recommendation and delivery. This assumption is considered reasonable for indoor or pedestrian users.

When a user generates a request according to the received recommendation list, we assumed that the probability of requesting any content in the recommended list is the same. When a specific piece of content is requested, traffic demand with respect to the content size is generated. As shown in Figure 2, different content sizes will result in varying traffic demands across different time slots. The probability mass function (PMF) of the traffic demand from user u at time slot t is given by

$$Q_u^t(q) = \begin{cases} 1-\lambda_u, & q = 0, \\ \sum_{f \in L_u} \frac{\lambda_u}{N} \delta(q - L_{uf}), & \text{otherwise}, \end{cases} \tag{2}$$

or expressed as

$$Q_u^t(q) = (1-\lambda_u)\delta(q) + \sum_{f \in L_u} \frac{\lambda_u}{N} \delta(q - L_{uf}), \tag{3}$$

where N is the total number of files in the recommendation list, L_u is the set of indexes of the subset of contents included in the u-th user's recommendation list, and L_{uf} is the size of the corresponding file in the u-th user's recommendation list. Because we have a fixed number of files, the density of traffic demand at a time slot can only take discrete values of the file sizes.

When there are multiple users, then PMFs of each user's traffic at time slot t are given by

$$\begin{array}{c} Q_1^t(q) = (1-\lambda_1)\delta(q) + \sum_{f \in L_1} \frac{\lambda_1}{N} \delta(q - L_{1f}) \\ Q_2^t(q) = (1-\lambda_2)\delta(q) + \sum_{f \in L_2} \frac{\lambda_2}{N} \delta(q - L_{2f}) \\ \vdots \\ Q_U^t(q) = (1-\lambda_U)\delta(q) + \sum_{f \in L_U} \frac{\lambda_U}{N} \delta(q - L_{Uf}). \end{array} \tag{4}$$

The PMF of the sum traffic $Q^t = q_1^t + q_2^t + \cdots + q_U^t = \sum_{u=1}^{U} q_u^t$ at time slot t is then

$$Q^t(q) = Q_1^t(q) \otimes Q_2^t(q) \otimes \cdots \otimes Q_U^t(q), \tag{5}$$

where $\otimes$ represents the convolution operation.

2.3. Air Interface Model

The air interface is in charge of signal modulation and transmission over the wireless links. It is a common bottleneck that limits the communication capacity of each user. Our paper considers an orthogonal frequency division multiplexing (OFDM)-based air interface with K subcarriers. Each subcarrier is an independent channel that adapts M-quadrature amplitude modulation (M-QAM). The subcarrier index is denoted by $k (k = 1, 2, \ldots, K)$.

Let us denote c_{uk}, α_{uk}^2, and p_{uk} as the assigned bits of user u on subcarrier k, the instantaneous channel gain of user u on subcarrier k, and the transmission energy allocated to user u's subcarrier k, respectively. We have $p_{uk} = f(c_{uk})/\alpha_{uk}^2$, where $f(c)$ denotes the required received energy in the subcarrier for a reliable reception of c bits per symbol when the channel gain is equal to unity. In addition, we assume that the channel gains are exponentially distributed [25–27].

The total transmit power and bandwidth of the BS are denoted as P_T and B, respectively. In the presence of multiple users, the air interface should properly allocate the subcarrier (i.e., band width) and power resource to multiple users. The subcarrier allocation is indicated by a binary variable $\rho_{uk} \in \{0, 1\}$. If $\rho_{uk} = 1$, it means that subcarrier k is allocated to user u. Each subcarrier can only be allocated to at most one user.

Now, considering the transmission in a single subcarrier, the required received power $f(c)$ can be written as a function of the target BER P_e and bits per symbol c as [20,28]

$$f(c) = \frac{N_0}{3}\left[Q^{-1}\left(\frac{P_e}{4}\right)\right]^2 (2^c - 1), \tag{6}$$

where N_0 is the power spectrum density of white Gaussian noise and $Q^{-1}(\cdot)$ is the inverse Q-function [29]. For the u-th user, the bits per symbol duration aggregated over multiple subcarriers is given by

$$C_u = \sum_{k=1}^{K} c_{uk} \times \rho_{uk} \qquad \text{(bits/symbol)}. \tag{7}$$

Similarly, the sum energy per symbol duration for user u across multiple subcarriers is

$$P_u = \sum_{k=1}^{K} p_{uk} \times \rho_{uk} = \sum_{k=1}^{K} \frac{f(c_{uk})}{\alpha_{uk}^2} \times \rho_{uk}. \tag{8}$$

Because the symbol rate scales linearly with the system bandwidth B, the total power (per second) allocated to user u can be written as

$$P_u^{total} = B \times \sum_{k=1}^{K} p_{uk} \times \rho_{uk}. \tag{9}$$

Similarly, the bit rate (i.e., bits per second) of user u is

$$R_u = C_u \times B = \sum_{k=1}^{K} c_{uk} \times \rho_{uk} \times B \qquad \text{(bits/s)}. \tag{10}$$

2.4. Problem Formulation

In this paper, we are interested in the problem of UiL contert delivery. Upon a user request, a decision should be made to recommend a list of N contents to the user. Such a decision should jointly consider the user interest and the current channel conditions of the user. The decision could be made at the BSs in a distributed fashion, or at the central server in a centralized fashion. In both cases, the users' interest profile and channel conditions should be periodically monitored.

The performance of the UiL content delivery system can be evaluated by two metrics, both directly related to the user QoE. The first metric is sum user interests, which indicates the user satisfaction at the semantic level, while the second metric is transmission delay, which indicates the user satisfaction during the communication process. We assume that multiple users have different service classes and different delay requirements. The delay requirement of user u is denoted as τ_u. The recommendation algorithm needs to ensure that the contents requested by a user are delivered in time with high probability. Our problem formation aims to maximize the total user interests, under the condition that the probability for the transmission delay to be smaller than the given delay requirement is greater than a predefined parameter ξ. In other words, the probability of satisfying each user's delay requirement should be greater than ξ. Mathematically, the delay requirement can be written as

$$\forall u, t, P\left(\frac{q_u^t}{R_u^t} = \frac{q_u^t}{\sum_{k=1}^{K} c_{uk}^t \rho_{uk}^t \times B} \leq \tau_u\right) > \xi, \tag{11}$$

where q_u^t represents the amount of traffic generated by user u at time slot t, and R_u^t is the user's data rate at time slot t. The problem of UiL content delivery can be formally stated as

$$\begin{aligned}
\max \quad & \sum_{u=1}^{U}\sum_{f=1}^{F} r_{uf}x_{uf}, \\
\text{s.t.} \quad & \forall u,t, P(\frac{q_u^t}{R_u^t} = \frac{q_u^t}{\sum_{k=1}^{K} c_{uk}^t \rho_{uk}^t \times B} \leq \tau_u) > \xi, \\
& \forall t, B \times \sum_{u=1}^{U}\sum_{k=1}^{K} \frac{f(c_{uk}^t)}{\alpha_{uk}^2} \times \rho_{uk}^t \leq P_T, \\
& \sum_{f=1}^{F} x_{uf} = N, \\
& \sum_{u=1}^{U} \rho_{uk}^t \leq 1, \\
& x_{uf} \in \{0,1\}, \rho_{uk}^t \in \{0,1\}.
\end{aligned} \tag{12}$$

The goal of our problem is to optimize the total interest over multiple decision variables x_{uf}, ρ_{uk}^t and c_{uk}^t. This implies that the tasks of content recommendation and content delivery (i.e., subcarrier and power allocation) are jointly executed. Here, constraints C1 to C4 correspond to transmission delay constraint, total BS power limit, number of recommended files, and orthogonal subcarrier allocation, respectively.

3. Theoretical Performance Limits

Some theoretical performance bounds of the UiL content delivery system will be evaluated in this section. For convenience and mathematical tractability, we consider a simplified scenario with relatively ideal assumptions. Specifically, we assume that users' interest profiles are independent from each other. The users' interest coefficient r_{uf} follows independent and identical distributions (i.i.d.) given by a truncated normal distribution in $[a, a+h]$ with mean μ and variance δ^2. In addition, we suppose that content file sizes L_f follows a uniform distribution in $[b, b+g]$. Our subsequent analysis will try to derive the upper and lower bounds of the mean user interest.

3.1. Upper Bound

Given that the users' interest coefficient follows a normal distribution, the sum interest of the N contents with the highest interest for a user is

$$S = Z_{(F+1-N)} + Z_{(F+2-N)} + \ldots + Z_{(F),} \tag{13}$$

where $Z_{(i)}$ denotes the i-th order statistics. Referring to Refs. [30,31], the PDF of $Z_{(i)}$ can be evaluated by the order statistic theory as

$$f_i(z) = \frac{n!}{(i-1)!(n-i)!}(F(z))^{i-1}(1-F(z))^{n-i}f(z), z \in \Re, \tag{14}$$

where $f(z)$ and $F(z)$ are the probability density functions (PDFs) and cumulative distribution functions (CDFs) of the truncated normal distribution, respectively. The expectation of the i-th interest is

$$\mathbb{E}(i,n) = \int_a^b z f_i(z)\,\mathrm{d}z = \frac{n!}{(i-1)!(n-i)!}\int_a^b z(F(z))^{i-1}(1-F(z))^{n-i}f(z)\,\mathrm{d}z. \tag{15}$$

Then, we could get the mean of S as

$$\begin{aligned}\mathbb{E}(S;F;N) &= \mathbb{E}(Z_{(F+1-N)} + Z_{(F+2-N)} + \ldots + Z_{(F)}) \\ &= \mathbb{E}(Z_{(F+1-N)}) + \mathbb{E}(Z_{(F+2-N)}) + \ldots + \mathbb{E}(Z_{(F)}) \\ &= \mathbb{E}(F+1-N,F) + \mathbb{E}(F+2-N,F) + \cdots + \mathbb{E}(F,F).\end{aligned} \tag{16}$$

Here, $\mathbb{E}(S;F;N)$ denotes the sum interest in an individual user's recommended list. When we consider multiple users, the sum interest becomes

$$I_{total}^{est} = U \times \mathbb{E}(S;F;N). \tag{17}$$

3.2. Lower Bound

The lower bound of mean user interest lies in the case where the actual transmission delay of each user is required to be strictly lower or equal to τ, which means that ξ should be equal to 1. We have assumed that the file sizes of contents stored in the server obey a uniform distribution in $[b, b+g]$. If the data rate of user u is R_u, then the file size in each user's recommended list should be no more than $R_u * \tau$. We can then define a ratio as

$$\phi = \frac{R_u\tau - b}{g}. \tag{18}$$

Supposing that the number of available contents is proportional to the interval in uniform distribution, then the number of available contents under the above hypothesis is

$$\widehat{F} = F \times \phi. \tag{19}$$

With the number of contents given by $\widehat{F}$, the lower bound of mean user interest where the outage is zero can be obtained as

$$I_{total}^{0} = U \times \mathbb{E}(S;\widehat{F};N). \tag{20}$$

3.3. Simulation Validation

This subsection conducts Monte Carlo simulations to verify the theoretical performance bounds obtained above. The parameter values are set as follows. The BS power limit is $P_T = 2W$ while the number of channels $K = 256$. The user interest obeys a truncated normal distribution in $[1, 10]$ with mean 6 and variance 1. The file size of contents in the server obeys a uniform distribution in $[1, 50]$, the user behavior follows a Bernoulli process, the number of users is $U = 10$, the amount of content is $F = 500$, and the length of each user's recommended list is $N = 50$.

Figure 3 shows the total user interest as a function of transmission delay with different outage probability threshold ξ. The simulation performance curves are obtained by running the above-proposed algorithms for 100 random snapshots and taking the average over the results. In addition, the theoretical upper and lower bounds are also calculated and shown. According to Equation (17), the upper bound of total user interest is calculated to be 3874.7, which is not a function of the transmission delay. The lower bound of the total user interest is calculated according to Equation (20) and shown to be a function of the transmission delay. We can see that the theoretical upper and lower bounds depict a feasible region, which is an accurate description of the performance areas of the UiL content delivery system.

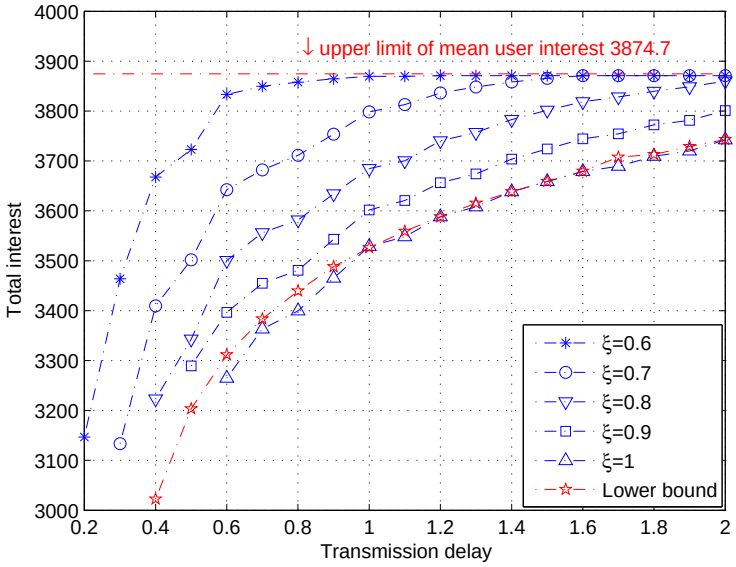

Figure 3. Theoretical bounds and simulation results of the total user interest as a function of the transmission delay with varying outage probability ξ.

4. Heuristic Algorithms for UiL Content Delivery Systems

Formulated in Equation (12), the problem of UiL content delivery is a nonlinear mixed integer programming problem. It can be easily shown that the problem is non-deterministic polynomial (i.e., NP-hard), such that the optimal solution is hard to obtain. Therefore, we retreat to a heuristic algorithm, which divides the problem into two phases: a radio resource allocation phase and a content recommendation phase. Optimized output of the first phase is used as the input of the second phase.

4.1. Capacity Allocation with Multiple Users

A commonly adopted heuristic for radio resource allocation is to maximize the sum capacity. This, however, is not suitable for our problem setting because different users have different delay requirements. To this end, we propose a refined heuristic by letting each user's capacity be inversely proportional to his/her delay constraint. In this case, the problem of radio resource allocation can be formulated as

$$\begin{aligned} \max \quad & \sum_{u=1}^{U}\sum_{k=1}^{K} c_{uk}^{t}\rho_{uk}^{t}, \\ \text{s.t.} \quad & \forall t, B \times \sum_{u=1}^{U}\sum_{k=1}^{K} \frac{f(c_{uk}^{t})}{\alpha_{uk}^{2}} \times \rho_{uk}^{t} \leq P_{T}, \\ & \sum_{u=1}^{U} \rho_{uk}^{t} \leq 1, \\ & R_{1}^{t} : R_{2}^{t} : \cdots : R_{U}^{t} = \frac{1}{\tau_{1}} : \frac{1}{\tau_{2}} : \cdots : \frac{1}{\tau_{U}}, \\ & \rho_{uk}^{t} \in \{0,1\}, \end{aligned} \tag{21}$$

where $R_{u}^{t} = \sum_{k=1}^{K} c_{uk}^{t}\rho_{uk}^{t} * B$, which denotes the data rate of user u at time slot t, and τ_{u} is the required transmission delay for user u. This problem is an integer programming problem, which is itself NP-hard. To solve this problem effectively, a greedy algorithm is proposed. The rationale of the algorithm is similar to the water-filling algorithm. The algorithm runs iteratively. In each iteration, a single bit is allocated to a user in a subcarrier. The allocation that demands the least power is chosen. The same process is repeated until the BS power constraint is reached.

The flowchart of the algorithm is shown in Figure 4. We note that our proposed algorithm is computationally efficient with a polynomial complexity. More specifically, the time complexity is $O(K^2)$ in the carrier allocation step and $O(P_T UK)$ in the bit allocation step.

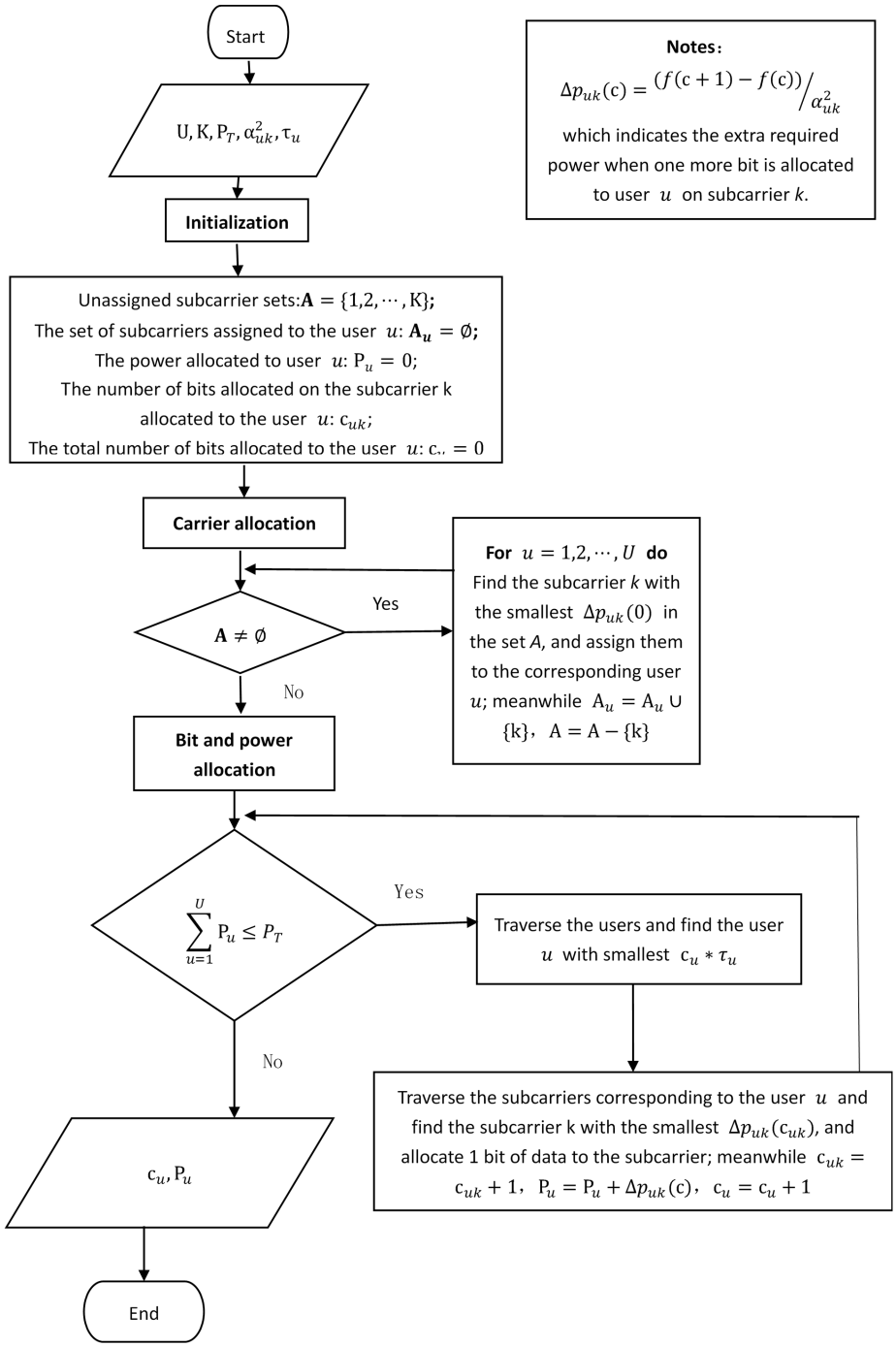

Figure 4. Proposed heuristic algorithm for capacity allocation with multiple users.

4.2. Recommendation of Content Lists

Once the capacity is assigned for each user, the problem of content recommendation can be formulated as

$$\begin{aligned} \max \quad & \sum_{u=1}^{U}\sum_{f=1}^{F} r_{uf}x_{uf}, \\ \text{s.t.} \quad & \forall u,t, P(\frac{q_u^t}{R_u^t} \le \tau_u) > \xi, \\ & \sum_{f=1}^{F} x_{uf} = N, \\ & x_{uf} \in \{0,1\}. \end{aligned} \tag{22}$$

This problem is still a nonlinear integer programming problem. Therefore, a novel heuristic algorithm is proposed to solve the problem in four steps: (1) Sort the user's interest profile in a descending order, then select the top N contents of each user as the initial solution, and calculate the total interest; (2) If each user's transmission delay constraint is satisfied, exit the algorithm. If not, continue to the next step; (3) For the users whose transmission delay constraint is not satisfied, discard the content with the largest file size in the recommendation list, and select the file whose size is smaller than the content just discarded and whose interest level is next in the ranking list; (4) Finally, recalculate the total interest and go back to step 2 to continue the calculation. The flowchart of the algorithm is shown in Figure 5. It can be shown that the algorithm has a polynomial time complexity given by $O(UFN + UF^2)$.

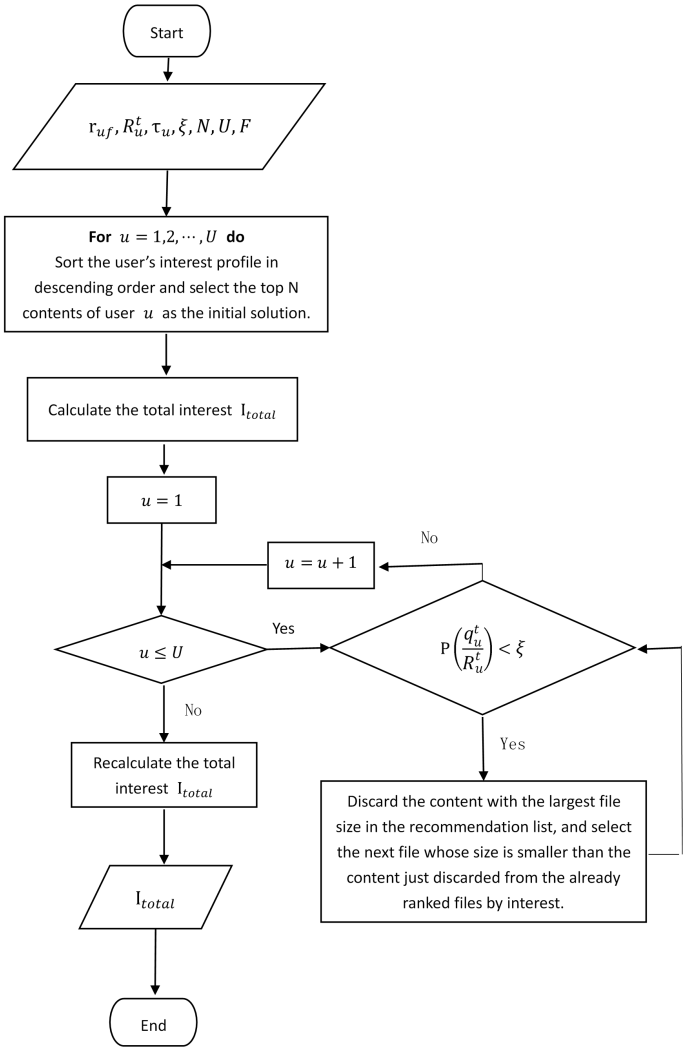

Figure 5. Proposed heuristic algorithm for content recommendation.

5. Simulation Results and Performance Evaluation

5.1. Realistic Models and Parameters

This section will give a comprehensive assessment on the performance of the proposed algorithm in realistic scenarios. Unlike Section 3 where ideal models are assumed, this section applies realistic models based on measurement. In the literature, the file size can follow either lognormal distributions [32–36] or power law distributions [37]. Here, we adapt the lognormal distribution because it is more frequently reported. The content popularity, which measures the aggregated interest of a particular content over multiple users, was reported to follow Zipf distribution [38] or power law distribution [39]. The widely-used Zipf distribution is adopted. Given a piece of content, the interest distribution among multiple users varies. It can be normal distribution [40], Levy alpha-stable distribution [41], Beta distribution [42], or U-shaped (or J-shaped) distribution [43,44]. The normal distribution is adopted in our paper by default. For the convenience of readers, Table 1 summarizes the parameters used in this section.

Table 1. Simulation parameters.

Simulation Parameter	Parameter Value
Number of users U	10
Number of contents F	500
Number of channels K	256
Recommended form length N	50
System bandwidth B	10 MHz
Noise power spectral density N_0	-174 dBm/Hz [45]
Bit error rate BER	1×10^{-4} [46]
Macrocell path loss model	$128.1 + 37.6\log_{10}d$ (d in km) [45]
Inter-site distance d	330 m
Channel gain α_{uk}^2	Exponential distribution of parameter 1
File distribution L	Logarithmic normal distribution with location parameter of 9.357 and scale parameter of 1.318 [33]
Interest matrix $\mathbf{R}$	Zipf distribution with parameter 1 [38]; Truncated Gaussian distribution between 1 and 5 with a mean of 3 and a variance of 2 [40].

5.2. Simulations Results and Discussion

The system performance is evaluated by the trade-off relationship between the total user interest and transmission delay. First of all, Figure 6 compares the performance between our proposed algorithm and two standard meta heuristic algorithms: the simulated annealing (SA) and genetic algorithm (GA) [47,48]. The simulated annealing algorithm can find a high-quality solution that does not strongly depend on the choice of the initial solution. It has been theoretically proved to converge to the optimum solution as long as the cooling process is slow enough [49]. On the other hand, the basic rationale of the genetic algorithm is to maintain a population of solutions to the problem and select the next generation solution according to the principle of survival of the fittest. The specific steps for SA and GA could be found in [50]. The simulation settings used in our paper for SA and GA are summarized in Table 2. It is shown in Figure 6 that our proposed algorithm outperforms the SA and GA algorithms, especially when the transmission delay is small.

In addition, for comparison purposes, we also show the performance of the traditional content delivery protocol based on, e.g., the "pull-type" http protocol. Because the traditional protocol does not include the "user-in-the-loop" procedure, it yields a static performance regardless of the delay

constraint. The static performance, shown as a single star sign in Figure 6, represents the case where the transmission delay is relaxed to an extreme to obtain the highest user interest. The above comparisons show that our heuristic algorithm not only offers greater trade-off flexibility than traditional protocols, but also outperforms standard meta-heuristic algorithms in solving the complicated NP-hard problem.

Table 2. Simulation settings for simulated annealing (SA) and genetic algorithm (GA).

Simulation Parameter	Parameter Value
Initial temperature T_0	0.5
Cooling method	$T_{i+1} = \beta T_i, \beta = 0.9$
Maximum number of iterations	500
Population size	30
Crossover percentage	0.5
Mutation percentage	0.4

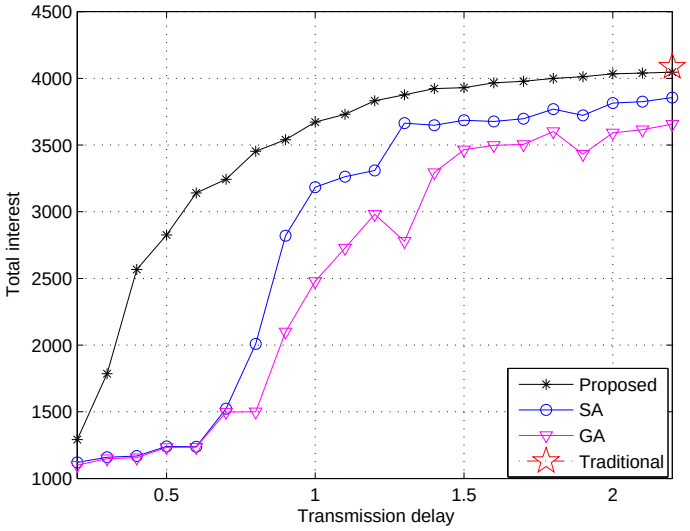

Figure 6. The total user interest as a function of the transmission delay with different algorithms ($P_T = 2W$, $\xi = 0.9$ and $\lambda = 0.5$).

Figure 7 investigates the impact of arrival rate λ on the system performance. The arrival rate represents the probability for a user to request a file from the central server at a time slot. It can be observed in Figure 7 that a lower arrival rate leads to a higher user interest. This is because a lower arrival rate means less competition for the limited radio resource, so that each user can be allocated with a higher capacity. It is interesting to see that a four times increase on the arrival rate (from 0.2 to 0.8) only causes less than 15% percent of reduction on the total interest. This means our system is robust to traffic congestion, such that it can cope well with increasing traffic demand per user or increasing number of users. We note that such a robustness essentially comes from our joint content recommendation and delivery design: when the per user capacity is reduced, content recommendation can choose to recommend small-size files with high user interests, so that the traffic demand can be effectively reduced while maintaining a high user interest level.

In Figure 8, we illustrate the total interest as a function of the transmission delay with varying values of the outage probability parameter ξ, which is the probability for the actual transmission delay to be less or equal to the required transmission delay. A greater value of ξ means a stricter requirement on the delay performance. It is observed that an increasing ξ leads to a reduced user interest. This is because a higher ξ implies that there is less tolerance to random delay outage. In addition, it is interesting to see that a high user interest with a small transmission delay of 0.4 could be obtained

when the outage probability parameter $\xi = 0.6$. This implies that, in practice, a system designer should aim to strike a balance by jointly choosing proper values for ξ and the transmission delay.

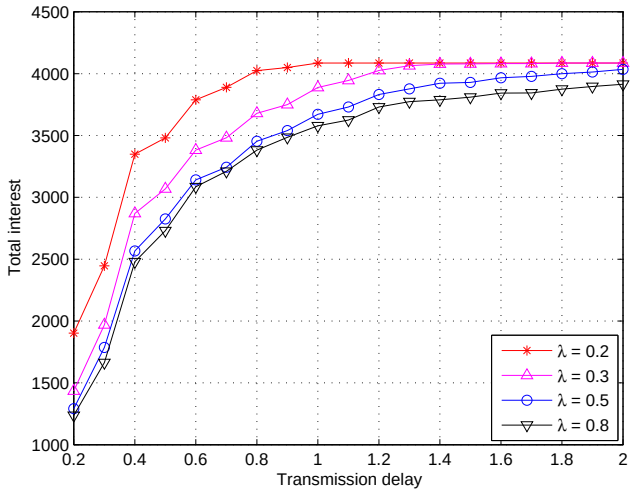

Figure 7. The total user interest as a function of the transmission delay with varying user arrival rate λ ($P_T = 2W$ and $\xi = 0.9$).

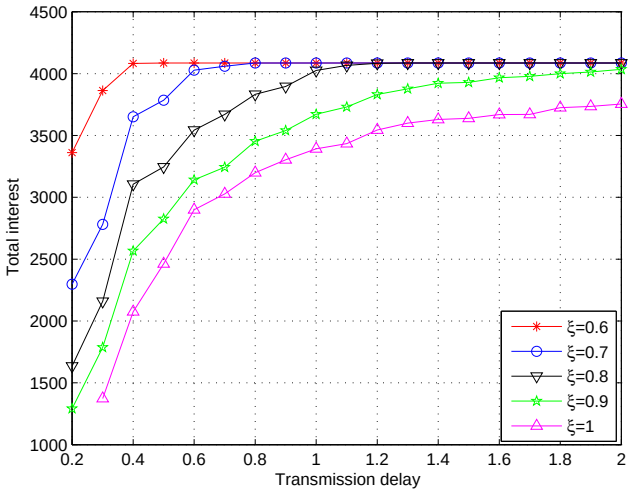

Figure 8. The total user interest as a function of the transmission delay with varying outage probability ξ ($P_T = 2W$ and $\lambda = 0.5$).

Furthermore, Figure 9 demonstrates the impact of the BS power constraint on the trade-off performance. Intuitively, the overall user interest is observed to increase with increasing power. When the power is large, an ideal solution with low transmission delay and high user interest could be achieved. However, increasing the power supply indefinitely will have diminishing returns in terms of energy efficiency. We further note that the aspect of intercell interference is not addressed in this paper. In practice, high BS power will drive the capacity to the interference-limited region, where the energy efficiency can be arbitrarily low. In practice, the power constraint should be properly chosen according to a targeted interest–delay trade-off performance.

Finally, Figure 10 demonstrates how the proposed system can be used to effectively differentiate users' QoE/QoS. As shown in Equation (21), our system supports customized user QoS by allowing different users to have different delay constraints. The proposed radio resource allocation algorithm allocates a user's data rate to be inversely proportional to its delay. As a result, high-priority users with low transmission delay will be allocated with more radio resources and will hence enjoy better user

QoS/QoE. A proper measure of the QoE is the individual user's interest (instead of the sum interest). Figure 10 shows the individual user's interest as a function of the traffic arrival rate λ with varying user delay constraints. It is observed that the user interests can be effectively differentiated by setting different delay constraints τ_u.

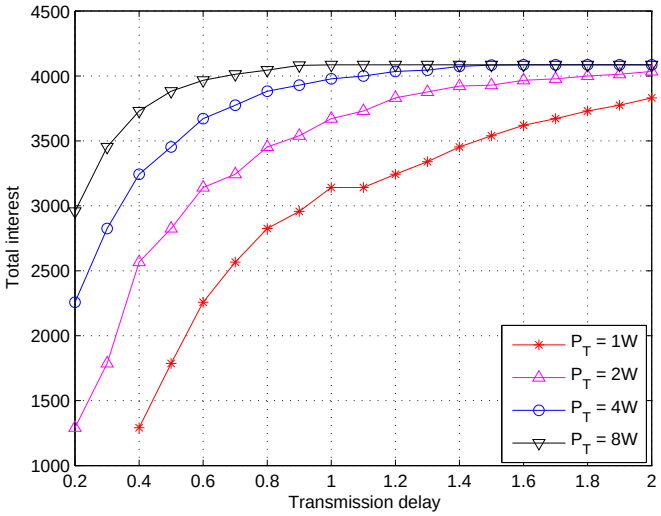

Figure 9. The total user interest as a function of the transmission delay with varying power constraint P_T ($\lambda = 0.5$ and $\xi = 0.9$).

In addition, Figure 10 also evaluates the joint impacts of traffic density λ and outage probability threshold ξ by varying both parameters together. Intuitively, it is observed that the user interest decreases with increasing density and outage threshold. However, it is interesting to see that, when the outage threshold is relatively low (i.e., more tolerance on outage), the user interests remain steady with increasing traffic density until λ reaches a certain threshold. In other words, a smaller outage threshold corresponding to a higher traffic density threshold, below which changing the traffic density has a marginal impact on the user interest. As shown in Figure 10, the outage thresholds at 0.6 and 0.9 yields density thresholds at 0.5 and 0, respectively.

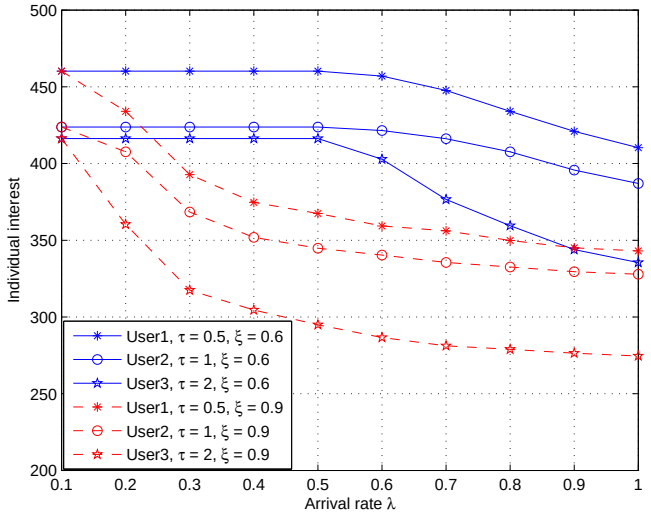

Figure 10. The individual user interest as a function of the arrival rate with varying user delay and outage requirements ($U = 3$ and $P_T = 0.5W$).

6. Conclusions

In this paper, we have investigated the problem of UiL content delivery over mobile communication systems. The problem has been formulated under a generalized model with asynchronous user behavior and customized user QoS requirements. The upper and lower performance bounds of the system have been theoretically derived in simplified scenarios. For more realistic scenarios, a novel heuristic algorithm with polynomial complexity has been presented and shown to outperform conventional schemes and meta-heuristic algorithms. Simulation results have shown that by exploiting the user interest semantics as a new degree of freedom for system design, the UiL system offers a holistic approach to better manage the overall QoE of personalized content retrieval services.

Author Contributions: W.Z., X.H. and J.S. conceived and designed the experiments; Y.L. and H.L. performed the experiments; Y.L. and W.Z. analyzed the data; W.Z., Y.L. and X.H. wrote the paper. All authors have read and approved the final manuscript.

Acknowledgments: The authors acknowledge the support from the Natural Science Foundation of China (Grant No. 61571378).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ekstrand, M.D.; Riedl, J.T.; Konstan, J.A. Collaborative filtering recommender systems. *Found. Trends Hum. Comput. Interact.* **2011**, *4*, 81–173. [CrossRef]
2. Ricci, F.; Rokach, L.; Shapira, B. Recommender systems: introduction and challenges. In *Recommender Systems Handbook*; Springer: New York, NY, USA, 2015; pp. 1–34.
3. Yao, L.; Sheng, Q.Z.; Wang, X.; Zhang, E.W.; Qin, Y. Collaborative Location Recommendation by Integrating Multi-dimensional Contextual Information. *ACM Trans. Internet Technol.* **2017**, *18*. [CrossRef]
4. Wang, X.; Zhao, Y.L.; Nie, L.; Gao, Y.; Nie, W.; Zha, Z.J.; Chua, T.S. Semantic-based location recommendation with multimodal venue semantics. *IEEE Trans. Multimedia* **2015**, *17*, 409–419. [CrossRef]
5. Antonopoulos, A.; Kartsakli, E.; Perillo, C.; Verikoukis, C. Shedding Light on the Internet: Stakeholders and Network Neutrality. *IEEE Commun. Mag.* **2017**, *55*, 216–223. [CrossRef]
6. Wang, Z.; Zhu, W.; Chen, M. CPCDN: Content delivery powered by context and user intelligence. *IEEE Trans. Multimedia* **2015**, *17*, 92–103. [CrossRef]
7. Lum, W.Y.; Lau, F.C.M. A context-aware decision engine for content adaptation. *IEEE Pervasive Comput.* **2002**, *1*, 41–49.
8. Shoukry, O.; ElMohsen, M.A.; Tadrous, J. Proactive scheduling for content pre-fetching in mobile networks. In Proceedings of the 2014 IEEE International Conference on Communications, Sydney, Australia, 10–14 June 2014; pp. 2848–2854.
9. Shoukry, O.K.; Fayek, M.B. Evolutionary scheduler for content pre-fetching in mobile networks. In Proceedings of the 2013 AAAI Fall Symposium Series, Arlington, VA, USA, 15–17 November 2013; pp. 386–391.
10. Tadrous, J.; Eryilmaz, A.; Gamal, H.E. Proactive content download and user demand shaping for data networks. *IEEE/ACM Trans. Netw.* **2015**, *23*, 1917–1930. [CrossRef]
11. Weng, X.; Baras, J.S. Joint optimization for social content delivery in wireless networks. In Proceedings of the 2016 IEEE International Conference on Communications, Kuala Lumpur, Malaysia, 23–27 May 2016; pp. 1–7.
12. Weng, X.; Baras, J.S. Joint optimization for social content delivery in heterogeneous wireless networks. In Proceedings of the 2016 14th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks, Tempe, AZ, USA, 9–13 May 2016; pp. 1–8.
13. Yin, L.; Cao, G. Adaptive power-aware prefetch in wireless networks. *IEEE Trans. Wirel. Commun.* **2004**, *3*, 1648–1658. [CrossRef]
14. Gungor, A.C.; Gunduz, D. Proactive wireless caching at mobile user devices for energy efficiency. In Proceedings of the 2015 International Symposium on Wireless Communication Systems, Brussels, Belgium, 25–28 August 2015; pp. 186–190.

15. Tadrous, J.; Eryilmaz, A.; El Gamal, H. Joint smart pricing and proactive content caching for mobile services. *IEEE/ACM Trans. Netw.* **2016**, *24*, 2357–2371. [CrossRef]
16. Tadrous, J.; Eryilmaz, A.; El Gamal, H. Pricing for demand shaping and proactive download in smart data networks. In Proceedings of the 2013 IEEE Conference on Computer Communications Workshops, Turin, Italy, 14–19 April 2013; pp. 321–326.
17. Giatsoglou, N.; Ntontin, K.; Kartsakli, E.; Antonopoulos, A.; Verikoukis, C. D2D-Aware device caching in mmWave-cellular networks. *IEEE J. Sel. Areas Commun.* **2017**, *35*, 2025–2037. [CrossRef]
18. Antonopoulos, A; Kartsakli, E; Verikoukis, C. Game theoretic D2D content dissemination in 4G cellular networks. *IEEE Commun. Mag.* **2014**, *52*, 125–132. [CrossRef]
19. Ji, M.; Caire, G.; Molisch, A.F. Wireless device-to-device caching networks: Basic principles and system performance. *IEEE J. Sel. Areas Commun.* **2016**, *34*, 176–189. [CrossRef]
20. Li, Y.; Chen, L.; Shi, H.; Hong, X.; Shi, J. Joint Content Recommendation and Delivery in Mobile Wireless Networks with Outage Management. *Entropy* **2018**, *20*, 64. [CrossRef]
21. Bobadilla, J.; Ortega, F.; Hernando, A. Recommender systems survey. *Knowl. Based Syst.* **2013**, *46*, 109–132. [CrossRef]
22. Yang, X.; Guo, Y.; Liu, Y. A survey of collaborative filtering based social recommender systems. *Comput. Commun.* **2014**, *41*, 1–10. [CrossRef]
23. Adomavicius, G.; Tuzhilin, A. Context-aware recommender systems. In *Recommender Systems Handbook*; Springer: New York, NY, USA, 2015; pp. 191–226.
24. Koren, Y.; Bell, R.; Volinsky, C. Matrix factorization techniques for recommender systems. *Computer* **2009**, *42*, 30–37. [CrossRef]
25. Goldsmith, A. *Wireless Communications*; Cambridge University Press: Cambridge, UK, 2005.
26. Hasna, M.O.; Alouini, M.S. End-to-end performance of transmission systems with relays over Rayleigh-fading channels. *IEEE Trans. Wirel. Commun.* **2003**, *2*, 1126–1131. [CrossRef]
27. Wang, Z.; Giannakis, G.B. A simple and general parameterization quantifying performance in fading channels. *IEEE Trans. Commun.* **2003**, *51*, 1389–1398. [CrossRef]
28. John, M.C. Signal Processing and Detection. Available online: http://web.stanford.edu/group/cioffi/ee379a/course_reader/chap1.pdf (accessed on 23 March 2018).
29. Kay, S.M. *Fundamentals of Statistical Signal Processing, Vol. II: Detection Theory*; Prentice Hall: Englewood Cliffs, NJ, USA, 1998.
30. Kyle, S. Random Samples. Available online: http://www.randomservices.org/random/sample/OrderStatistics.html (accessed on 23 March 2018).
31. Kamps, U. A concept of generalized order statistics. *J. Stat. Plan. Inference* **1995**, *48*, 1–23. [CrossRef]
32. Evans, K.M.; Kuenning, G.H. A study of irregularities in file-size distributions. In Proceedings of the 2002 International Symposium on Performance Evaluation of Computer and Telecommunication Systems, San Diego, CA, USA, 14–18 July 2002.
33. Barford, P.; Crovella, M. Generating representative web workloads for network and server performance evaluation. *ACM Sigmetrics Perform. Eval. Rev.* **1998**, *26*, 151–160. [CrossRef]
34. Chlebus, E.; Divgi, G. A versatile probability distribution for light and heavy tails of web file sizes. In Proceedings of the 2009 Wireless Communications and Networking Conference, Budapest, Hungary, 5–8 April 2009; pp. 1–7.
35. Douceur, J.R.; Bolosky, W.J. A large-scale study of file-system contents. *ACM Sigmetrics Perform. Eval. Rev.* **1999**, *27*, 59–70. [CrossRef]
36. Gros, C.; Kaczor, G.; Markovi, D. Neuropsychological constraints to human data production on a global scale. *Eur. Phys. J. B Condens. Matter Complex Syst.* **2012**, *85*, 1–5. [CrossRef]
37. Arlitt, M.F.; Williamson, C.L. Internet web servers: Workload characterization and performance implications. *IEEE/ACM Trans. Netw.* **1997**, *5*, 631–645. [CrossRef]
38. Cha, M.; Kwak, H.; Rodriguez, P. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement, San Diego, CA, USA, 24–26 October 2007; pp. 1–14.
39. Adamic, L.A. Zipf, Power-Laws, and Pareto—A Ranking Tutorial. Available online: http://www.labs.hp.com/research/idl/papers/ranking/ranking.html (accessed on 23 March 2018).

40. Hu, N.; Zhang, J.; Pavlou, P.A. Overcoming the J-shaped distribution of product reviews. *Commun. ACM* **2009**, *52*, 144–147. [CrossRef]
41. Lorenz, J. Universality in movie rating distributions. *Eur. Phys. J. B Condens. Matter Complex Syst.* **2009**, *71*, 251–258. [CrossRef]
42. Del Rio, M.B.; Cocho, G.; Naumis, G.G. Universality in the tail of musical note rank distribution. *Phys. A Stat. Mech. Appl.* **2008**, *387*, 5552–5560. [CrossRef]
43. Hu, N.; Pavlou, P.A.; Zhang, J. Can online reviews reveal a product's true quality? Empirical findings and analytical modeling of Online word-of-mouth communication. In Proceedings of the 7th ACM Conference on Electronic Commerce, Ann Arbor, MI, USA, 11–15 June 2006; pp. 324–330.
44. Cai, T.; Cai, H.J.; Zhang, Y. Polarized score distributions in music ratings and the emergence of popular artists. In Proceedings of the Science and Information Conference, London, UK, 7–9 October 2013; pp. 472–476.
45. LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Frequency (RF). Requirements for LTE Pico Node B. ETSI TR 136 931 v9.0.0. Available online: http://www.etsi.org/deliver/etsi_ts/136100_136199/136104/09.04.00_60/ts_136104v090400p.pdf (accessed on 23 March 2018).
46. Kim, I.; Park, I.S.; Lee, Y.H. Use of linear programming for dynamic subcarrier and bit allocation in multiuser OFDM. *IEEE Trans. Veh. Technol.* **2006**, *55*, 1195–1207. [CrossRef]
47. Blum, C.; Roli, A.; Alba E. An introduction to metaheuristic techniques. In *Parallel Metaheuristics: A New Class of Algorithms*; John Wiley & Sons, Inc.: New York, NY, USA, 2005; Volume 47.
48. Boussaid, I.; Lepagnot, J.; Siarry, P. A survey on optimization metaheuristics. *Inf. Sci.* **2013**, *237*, 82–117. [CrossRef]
49. Aarts, E.; Korst, J. *Simulated Annealing and Boltzman Machines: A Stochastic Approach to Combinatorial Optimization and Neural Computing*; Wiley: New York, NY, USA, 1989.
50. Khajehzadeh, M.; Taha, M.R.; El-Shafie, A.; Eslami, N. A survey on meta-heuristic global optimization algorithms. *Res. J. Appl. Sci. Eng. Technol.* **2011**, *3*, 569–578.