MediaServ: Subscription-based Media Crowdsourcing for Collective Events

A s m Rizvi, Shamir Ahmed, Minhajul Bashir, and Md. Yusuf Sarwar Uddin
Department of CSE, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh
Email: {rizvi22222, shamir253, shovan.livebd}@yahoo.com, yusufsarwar@cse.buet.ac.bd

Abstract—In this paper we propose resource optimization for subscription based media content crowdsourcing. In this form of crowdsourcing, interested entities (referred to as Campaigners) announce their interests expressing what media content they want to receive from users whereas users subscribe to those interests to serve content satisfying the interests. Campaigners solicit content generated by users by mentioning criteria that the media content should satisfy, for example, a noise pollution campaigner who wants to measure noise level of a city, may ask for audio clips recorded at a certain location. Subscribed users voluntarily or on paid terms generate content against those interests. Given that a user may subscribe to different campaign interests, its generated content may satisfy different interests in varying degree of accuracy, we propose methods to evaluate contents based on the degree of satisfaction against the interests, and develop techniques for delivering those contents.

I. INTRODUCTION

In this paper, we propose a media data collection system that collects media content from citizens against expressed subscriptions for collective events. We refer collective events as a set of social or citizen assisted events that require content contribution from participants. This paper proposes content upload policies for subscription based media crowdsourcing (SMCS). SMCS refers to a participatory platform where interested parties can create events and ask ordinary citizens to voluntarily deliver a certain type of media contents to a service point of the event generator. Example may include a City authority asking its citizens to report problems of inhabitants by taking photos, such as an open sanitary leak. Department of environment may ask citizens to capture audio clips for reporting noise level, or transportation department may ask for pollution incidents captured by videos of vehicles releasing black smokes. This type of service modality in emerging due to two reasons. One is the abundance of mobile phones equipped with high resolution camera, audio/video capture capabilities possessed by average individuals (study predicts the number of phones may exceed 2 billion in 2015 [1]). The second reason is the generation and distribution of huge amount of media contents through applications, such as Instagram, Flickr (photo sharing), YouTube, Vimeo (video sharing), and SoundCloud (audio sharing). Record shows nearly 350 million photos are uploaded daily on Facebook [2]. This paper leverages these opportunities to address a question whether we can channelize these contents in more targeted end points by explicit intents and generate values from the content uploaded. The initial idea of this paper was also appeared in [3]. In this paper, we elaborate the initial idea with simulation and real experiment results.

Subscription based media crowdsourcing is yet another publication, a set of requesters, we call them campaigners, solicit content from users by mentioning explicit interests of certain media content. These requests are referred to as campaigns, which are described by a set of conditions that the requested contents should satisfy (e.g., pictures taken from a certain location during daytime). These citizen assisted events need large amount of media contents from participants of users. Interested mobile users may register to these campaigns and publish content. Unlike traditional pub/sub where mobile user are mostly treated as subscribers receiving published contents, here in our context they are rather publishers, uploading contents for campaigns. As users may subscribe to several such campaigns or events, an issue becomes important how much resource the user may want to engage in to serve them. Admittedly media contents are large data objects so it usually takes considerable amount of resources to process and upload them. In this paper, we address the problem of content uploading in the context of subscription based media crowdsourcing for collective events.

Since mobile users rely on metered 3G or 4G connections that provide only limited data usage (as they are charged heavily otherwise), data transfer capacity becomes one of the sole constraints; more specifically the budget for content uploads to the service end points. Our formulation takes this into consideration. As user generated contents may satisfy different subscribed interests in varying degree of accuracy, we propose methods to evaluate those contents based on the degree of matching against the subscribed interests. Again, different services could have different level of importance (may be set forth by users). Considering all these, we develop a formulation for delivering contents that optimizes the user’s resource utilization.

To this end, we propose MediaServ, a system that handles subscription based media crowdsourcing for a mobile device and optimally allocates resources for multiple such subscriptions. Even though the proposed system is quite general in terms of types of media content it can serve (such as image, audio), we explicitly limit our work and experiments on pictures, mainly because pictures are predominant media content to date, and their use-cases are most befitting to our design. In that, we assume users subscribe to interests that ask for pictures. Our system has an Android based app that manages subscriptions, captures and uploads images against registered campaigns respecting individual data budget constraint.

II. RELATED WORKS

Some photo retrieval service also exists. One example is MediaScope [4], which tries to fill up the availability gap.
Mobile participatory sensing has also been used for capturing sound data. One such system has been proposed in [8], which focuses on measuring noise pollution levels. Another system, NoiseSPY [9], monitors noise to measure and annotate urban noise using smartphone sensing. NoiseTube [10] is another platform that utilizes noise data from smartphones as well as manual contextual data like location, time and noise source, and data generated by machine-based automatic classifiers.

In mobile participatory sensing, security and large user participation are the most common challenges. Many researches have been conducted regarding these two problems. Some incentive mechanism has been designed for mobile sensing in [11]–[13]. A recurrent reverse auction incentive mechanism with a greedy algorithm has been discussed in [11]. Based on Stackelberg game, and auction-based incentive mechanism a procedure is discussed in [12]. Efficient bidding system for user is discussed in [13]. Security issues are discussed in [14]–[16]. Framework Pepsi is discussed in [14]. Encrypted image based solution is given in [16]. Privacy-aware incentive mechanism is discussed in [17].

III. CONTENT UPLOAD IN SUBSCRIPTION BASED MEDIA CROWDSOURCING

Subscription based media crowdsourcing has three components: campaigners, interests, and publishers. Campaigners are individual persons or organizations who are in need of certain media content. Campaigners initiate a campaign by registering to the system by specifying their interests. These interests are simple predicates that media objects should satisfy (e.g., pictures from location X). Mobile users, act as publishers in the system, register to the campaigns they are interested to publish contents against. This registration is also known as subscription. Note that the term ‘subscription’ is meant different than its traditional meaning. Here subscription does not refer to receiving contents, rather publishing contents. When a content gets generated at a user (e.g., user takes a picture), the content is matched against all subscriptions whether it satisfies any interest of the subscribed campaigns of that user. Content is uploaded to the backend server only if it matches any subscription. Figure 1 shows the interaction.

While the development of entire pub/sub system is possible using existing pub/sub middleware, the full description with detail technical merit is beyond the scope of this paper. This paper focuses mainly on the operations that mobile publishers perform when they generate media contents and upload them to the system. We indeed do build a prototype system that entails all components, but may be in smaller scale, only enough to evaluate data publishing part of the system. Two important aspects are considered in this connection. One is the automatic scheduling of content uploads based on the subscriptions held by the user. The second one is related to specifying user’s data budget constraint on its content uploads. In the following, we describe these two issues.

A. Formulation of Content Uploads

When mobile users generate contents against their subscriptions, the contents are not uploaded immediately; instead they are uploaded in a batch, once in every so often defined by a period. This is for two reasons. Contents may arrive at a high rate in which immediate uploads may cause high processing at the mobile device. Again, the device may not be connected to the Internet allowing uploads not to happen anyway. We refer to these time instances as upload events when uploads can happen. If the user’s device is offline at the event time, the interval is stretched until the next connectivity occurs. Contents are queued between events.

Let \( c_1, c_2, \ldots, c_n \) be the set of contents (indexed by \( i \)) at an upload event, and \( s_1, s_2, \ldots, s_k \) be campaigns (indexed by \( j \)) that the user subscribed to. For each content, user may choose one or more campaigns where that content belong to. We define a matching score, \( \psi(i, j) \), that measures to what extend content \( i \) satisfies the requirement set forth by campaign \( j \). This is measured from media metadata and the requirement parameters set by the campaigners. Without of loss generality, we can assume \( 0 \leq \psi(i, j) \leq 1 \); where \( \psi(i, j) = 0 \), if content \( c_i \) is not assigned for \( s_j \) or it does not satisfy campaign requirement at all, and \( \psi(i, j) \) becomes 1 if it satisfies the interest completely. Any real number between 0 and 1 designates the degree of matching that content \( i \) demonstrates against the campaign requirement.

Campaigns may have differentiated importance or priorities overs others. These priorities can be set by the service providers when they register to the system or, alternatively, can be set by the user itself to differentiate among its subscribed campaigns. Priorities, denoted by \( p(j) \), are real numbers \( \geq 1 \), higher value means higher priority. Based on campaign priorities, we can define a value for each content as the weighted sum of all matching scores across all campaigns the

![Fig. 1: System component and flow](image-url)
content satisfies. That is, \( \text{value}(i) = \sum_j p(j)\psi(i,j) \). Note that \( \text{value}(i) = 0 \) means content does satisfy none.

Let \( x_i \) be a binary decision variable that indicates whether the content \( i \) should be uploaded at the current upload opportunity. Obviously, we discard items with \( \text{value}(i) = 0 \). Moreover, a user cannot upload all candidate objects at upload opportunities because it might have a data upload budget. We assume that users’ devices are on metered cellular connections and that users set explicit upload budget based on their connection plans. So, we want to upload contents so as to—

\[
\max \sum_i x_i \times \text{value}(i), \quad \text{subject to } \sum_i x_i \times \text{size}(i) \leq B
\]

where \( \text{size}(i) \) denotes the size (in byte) of content \( i \), and \( B \) is the total transfer volume allocated for at that upload event. Budget, \( B \), restricts how many bytes can be uploaded at this upload instance. The formulation is an instance of the classical 0/1 Knapsack problem which is reportedly NP-Hard. We use greedy heuristics solving for \( x_i \), particularly, pick items in descending order of \( \text{value}(i)/\text{size}(i) \) until the budget is full.

A few modifications can do the formulation is possible. We can limit low matching contents not to upload when their score falls below a threshold. The budget constraint can be written in various other forms too, such as

Budget per campaign

\[
\sum_i x_i \times \text{sz}(i) \leq B(j), \forall j
\]

Weighted budget

\[
\sum_i x_{ij} \times \text{sz}(i) \leq \left( \frac{p(j)}{\sum_i p(i)} \right) \times B, \forall j
\]

What remains to discuss is how to design matching function, \( \psi(i,j) \), for contents against campaigns. We do not tie this strongly in our service rather leave it open for campaigners to specify. For example, some campaigns may describe their interests by specifying explicit predicates over a set of attributes requesting items within a certain range of their attributes values (i.e., pictures from an area), which precisely defines the matching criteria. Another set of campaigns may specify a reference deviation from which measures degree of matching (i.e., pictures near city center) on Oct 31, 2015). In our system, we use spatio-temporal attributes, namely date, time, and location as reference attributes for requests, and define a distance function between a content and a campaign as follows:

\[
\text{dist}(i,j) = w_1\text{dist}_t(i,j) + w_2\text{dist}_l(i,j)
\]

where \( \text{dist}_t \) and \( \text{dist}_l \) measure much content \( i \) deviated in time and location for the reference point of campaign \( j \). Weights \( w_1 \) and \( w_2 \) define sensitivity to these two distance components. Afterwards, we simply use \( \psi(i,j) = e^{-D(i,j)} \) as the matching score. We assume each campaign specifies a spatio-temporal reference for their interests with an associated weight values.

**B. Data Budget Constraints for Mobile Users**

In our system mobile users use cellular data connections with limited data plan, which does not allow them to use all of their data plan for a single application. So, it is natural that they put a budget on their data activity per application. Therefore, our MediaServ app receives a budget, which is specified by a tuple \((D,T)\), indicating that the app is allowed to upload \( D \) amount of bytes for every \( T \) unit of time (we refer to as round). For simplicity, we assume \( T \) is an integral multiple of our scheduling period, \( P \). After \( T \) amount of time elapses, the user receives another chunk of \( D \) bytes to replenish its remaining budget for the successive operations. We assume that remaining budget rolls over from one round to the next.

MediaServ uses two budget allocation techniques to determine how much data can be uploaded at each upload events. Recall that upload events are at least \( P \) time apart. In the first scheme, the app assumes all \( D \) bytes are available from the beginning and uses up the budget as upload happens until it hits zero. This scheme is called gradual decrease. In the other scheme, budget is released rather slowly at a constant rate \( R = \frac{B}{T} \). That means, full \( D \) bytes are not available at the beginning; rather accumulates as time passes. This technique is referred to as gradual release. In this case, the application maintains a running budget counter, \( B \), to hold much it is allowed to spend now, which is updated as follows:

\[
B_{\text{now}} = B_{\text{old}} + R(t_{\text{now}} - t_{\text{last}}), \quad \text{where } t_{\text{now}} \text{ means current time and } t_{\text{last}} \text{ the last time a budget was replenished.}
\]

Even though both schemes are confirmed to use the same amount of uploads, we see in experiments that graduate release performs better than gradual decrease.

We also propose another budget allocation scheme that lies in between of the above two. This uses a varying replenishment rate in contrast to constant rate as in gradual release. The gradual release scheme releases take the full duration, \( T \), to release the entire budget at a constant rate, \( r(t) = R \). In varying rate case, we allow entire budget to get released by \( \beta T \) time for a control parameter \( \beta \leq 1 \). That means, at the beginning the process has a higher rate of release, and then the rate slows linearly down to zero after \( \beta T \). We can visualize the rate line, \( r(t) \), along time, \( t \), to form a right triangle whose base in \( \beta T \), and the height is such that the entire area bounded by the triangle is \( D \) (i.e., total budget). So, the height is \( \frac{2D}{\beta T} = \frac{2R}{\beta} \), which gives the rate equation as follows:

\[
r(t) = \frac{2R}{\beta} \left( 1 - \frac{t}{\beta T} \right)
\]

Hence, update rule for \( B \) for this scheme becomes:

\[
B_{\text{new}} = B_{\text{old}} + \int_{t_{\text{last}}}^{t_{\text{now}}} r(t) dt
\]

**C. Scheduling for Multiple Campaigns**

When a user subscribes to multiple campaigns, the formulation mentioned in Section III-A favors campaigns with higher priority, which may cause an unfair distribution of resources across campaigns. To mitigate this, in addition to greedy heuristic mentioned earlier, we propose a few heuristics that addresses usage of upload resources at varying degree of fairness.

First obvious choice could be FIFO order, contents are uploaded in temporal order. This requires little processing on content collection as it does not need further processing other than maintaining a queue. Another approach can be round robin; one content from each campaign at a time and continuing. In this, all campaigns are treated alike, and their
priority values are ignored. Contents under each campaign however sorted in descending order of their value-size ratios. This perhaps gives the most fair usage of upload opportunities across campaigns. A variant of round robin can be probabilistic method where campaigns are chosen randomly based on the weighted priority. In that, the probability that campaign \( j \) would be chosen next is equal to \( \frac{p(j)}{\sum_{i} p(i)} \). The final one can be strictly in order of campaign priority. Contents from the top campaigns are all uploaded first then moving to the next important campaign. This is useful in contexts where campaigns are very short-lived and required contents all in a certain time deadline (for example, journalists cover stories for tomorrow’s daily).

### IV. Simulation Results

We do a simulation based study to demonstrate our budget scheme on content uploads. We build a custom discrete event simulation using Java to simulate a mobile client with varying network connectivity (the device ON/OFF pattern is modeled using a Poisson process). We use a Poisson arrival process of media content generation with a rate of 1 per thirty minutes. Contents are images of random size 0.5MB to 2.5MB. We use varying degree of data budgets for a duration of 15 days.

To introduce variability in content generation and their impact on uploads against a given budget allocation scheme, we use three more data generation scenarios in addition to a constant rate. These incorporate various rates at different interval of the simulation. First one generates content at a high rate at the beginning during first half followed by average rate at the second half. The next is the reverse: low followed by high. The last one alternates between high and low for a quarter duration of each. Table 1 shows the simulation environment.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content value(priority)</td>
<td>1-10</td>
</tr>
<tr>
<td>Content size</td>
<td>0.5 MB to 2.5 MB</td>
</tr>
<tr>
<td>Content generation rate</td>
<td>1 per 30 minutes (Five times faster at high rate)</td>
</tr>
<tr>
<td>Network on</td>
<td>Average 10 minutes</td>
</tr>
<tr>
<td>Network off</td>
<td>Average 30 minutes</td>
</tr>
<tr>
<td>Simulation run for</td>
<td>15 days</td>
</tr>
<tr>
<td>Net speed</td>
<td>50 kbps</td>
</tr>
<tr>
<td>In gradual budget</td>
<td>Increase by a constant(budget/days) after every day</td>
</tr>
<tr>
<td>Total budget</td>
<td>50 MB to 400 MB</td>
</tr>
<tr>
<td>Seed values</td>
<td>2000 to 2012</td>
</tr>
</tbody>
</table>

In all cases, we are interested in different performance metrics, such as number of content uploaded, total value of the all sent contents, total bytes sent, and number of dropped contents. Due to space limit, we here produce results on total content uploaded, referring to [18] for detailed results. In the following, we show results for a single subscription, leaving the same for multiple subscriptions (for full results, see [18]).

In Figure 2, we show comparison between two budgeting schemes, gradual release and decrease, at different content generation rates. It is apparent from all four graphs that, budget less than 250MB, gradual release scheme shows better result. After 250MB, gradual decrease shows similar or better results. Before 250MB, resource is scarce and we need to select which contents we choose to upload. But after 250MB, of total budget we have a large amount of budget and we can use it without much concerning about our limitations. We may find an intersection point where two schemes have the same results. After that, most of the cases, gradual decrease method shows good result. In Figure 2c and Figure 2d, the difference between the two graphs is more prominent.

When we wait for contents with higher values and delay uploads, the content list gets bigger as contents are generated during the wait time. When we get new allocation, may be we wasted a lot of time in waiting, hence we can not transfer as many contents as we can with gradual decrease method. This is the case when budget is larger than around 250MB. But when budget is very low, chance is that we send bigger sized contents too much and wasted a lot of bandwidth. That is why from all the figures, we can see for lower amount of budget gradual release method shows better result.

In Figure 2b, we see that there is no significant difference between two curves. This is the case when content generation rate is very high at the beginning. Hence, there are many contents which were created during the high generation rate period. Gradual decrease method starts to send the contents immediately but gradual release method delays upload because it has low budget at the beginning. As the content list grows, the scheme gets the chance to find smaller sized contents among many contents. Hence, in this scenario it sends somehow more smaller sized contents than the other generation rate scenario. If most of the contents are generated within a short interval, then the gradual decrease method gives better result and the transmission of the contents will be much faster than the gradual release method.

Even though both schemes send varying number of contents, they effectively upload almost the same amount of bytes (see Figure 3a and Figure 3b), which indicates that graduate release uploaded more smaller sized objects than large sized one.

![Fig. 3: Comparison between budgeting methods at different rate of content generation](image-url)

Budget can be released in a different rate rather than a constant rate. The release rate may be higher than the constant rate. As a result all the budget will be released earlier, before the end of the simulation. We simulated for two different rates. A rate indicates the fraction of total time, that is necessary to release all the budget. In Figure 5a and Figure 5b we see the comparison. At the beginning we have two lines in between gradual release and gradual decrease lines. When rate is 0.5 the line follows the line of gradual decrease method. This is because the whole budget is released within \( \frac{1}{2} \) time. This is
(a) Constant rate
(b) High rate followed by low rate
(c) High low alternates
(d) Low rate followed by high rate

Fig. 2: Comparison between budgeting methods at different rates of content generation

some-what likely with gradual decrease method.

(a) No. of transferred content comparison
(b) Total sent value(priority) comparison

Fig. 4: Comparison among budgeting methods

One more thing, we can see in all graphs that the lines are capped after around 250MB budget. As we generate 1 content per 30 minutes, there will be around 720 contents in fifteen days. But before transferring all of these and fills up the whole budget the 15 days time of simulation gets finished. That is why the lines are capped after 250MB of budget.

We can conclude that gradual release budget scheme in general is better than its counterpart. Actually, the performance is largely dependent on content generation rate and their sizes. We observe that when the budget is small following gradual release method is good. When the budget is large gradual decrease shows some promise for better results.

We also did experiments with varying mean network ON time and varying data transfer rate (defaults are average 10 minutes and 50 kbps respectively). Network ON time indicates on an average how long the cellular connection remains once it gets connectet. Figure 5 shows results. We see that as we increase network ON time and transfer bandwidth more and more pictures are uploaded.

We tested the app for three campaigns with a small number of users:

• Clean campus campaign: This campaign asked students to report dirts/littered objects spread around in our campus. This is a awareness campaign against reckless littering of objects in a campus.
• Festival campaign: This asked students to take photos of events of a week long student program known
as CSE FESTIVAL 2015. This asks for photos from within campus as well as within a certain period of time.

- Spelling campaign: Shops in cities frequently misspell common words in their displays. The purpose of this campaign is to collect pictures of these spelling mistakes in a nearby market area.

The following table shows how many photos are taken per campaign and how many of them uploaded during the campaign duration (details in [18]).

<table>
<thead>
<tr>
<th>Campaign</th>
<th>Generated</th>
<th>Uploaded</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean campus campaign</td>
<td>31</td>
<td>21</td>
</tr>
<tr>
<td>Festival campaign</td>
<td>38</td>
<td>24</td>
</tr>
<tr>
<td>Spelling campaign</td>
<td>42</td>
<td>32</td>
</tr>
</tbody>
</table>

VI. CONCLUSION

In this paper, we mainly focus on service based image retrieval system. Evaluation technique of contents is illustrated. As users normally are not willing to give data in participatory network, hence a technique in where users can send content efficiently is illustrated in this text. Maximize the service campaign’s gain, and optimize the user’s resource is mainly described. Efficient way of image processing in user device can be added in future, which will make the content evaluation process more efficient. We also tried to make an effective manner of using user bandwidth with two budgeting schemes. Various scheduling methods are described which can be used in an adaptive manner in different circumstances.

REFERENCES