Who Should Google Scholar Update More Often?

Melih Bastopcu  
Department of Electrical and Computer Engineering  
University of Maryland, College Park, MD 20742  
bastopcu@umd.edu

Sennur Ulukus  
Department of Electrical and Computer Engineering  
University of Maryland, College Park, MD 20742  
ulukus@umd.edu

Abstract—We consider a resource-constrained updater, such as Google Scholar, which wishes to update the citation records of a group of researchers, who have different mean citation rates (and optionally, different importance coefficients), in such a way to keep the overall citation index as up to date as possible. The updater is resource-constrained and cannot update citations of all researchers all the time. In particular, it is subject to a total update rate constraint that it needs to distribute among individual researchers. We use a metric similar to the age of information: the long-term average difference between the actual citation numbers and the citation numbers according to the latest updates. We show that, in order to minimize this difference metric, the updater should allocate its total update capacity to researchers proportional to the square roots of their mean citation rates. That is, more prolific researchers should be updated more often, but there are diminishing returns due to the concavity of the square root function. More generally, our paper addresses the problem of optimal operation of a resource-constrained sampler that wishes to track multiple independent counting processes in a way that is as up to date as possible.

I. INTRODUCTION

Consider a citation index such as Google Scholar. As abstracted out in Fig. 1, Google crawls the web to find and index various items such as documents, images, videos, etc. Focusing on scientific documents, Google Scholar further examines the contents of these documents to extract out citation counts for indexed papers. Google Scholar then needs to update citation counts of individual researchers, which there are many. We model the citation count of each individual researcher as a counting process with a fixed mean, e.g., $\lambda_i$ for researcher $i$. Assuming that Google Scholar is resource-constrained, i.e., that it cannot update all researchers all the time, how should it prioritize updating researchers? If it can update only a fraction of all researchers, who should it update? Should it update researchers with higher mean citation rates more often as their citation counts are subject to larger change per unit time? Or should it update researchers with lower mean citation rates more often in order to capture rarer more informative changes?

We view this problem with the lens of the recent literature on information freshness quantified through the metric of age of information. freshness, and age, of information have been studied in the context of web crawling [1]–[4], social networks [5], queueing networks [6]–[15], caching systems [16], [17], remote estimation [18]–[21], energy harvesting systems [22]–[31], fading wireless channels [32], [33], scheduling in networks [34]–[41], multi-hop multicast networks [42]–[45], lossless and lossy source coding [46]–[50], computation-intensive systems [51]–[56], vehicular, IoT and UAV systems [57], [58], reinforcement learning [59]–[61] and so on.

We model the citation count of each researcher as a counting process with a given mean. In particular, we model citation arrivals for researcher $i$ as a Poisson counting process with rate $\lambda_i$. Optionally, we may further assign an importance factor to each researcher, based on their research field or citation count, but this is optional, and does not affect the structure of the results. If an importance coefficient is used, we denote it with $\mu_i$ for researcher $i$. Ideally the updater should update all researchers all the time, however, due to computational limitations, this may not be possible. We model the updater as a resource-constrained entity which has a total update capacity of $c$, which it should distribute among all researchers. We allocate an update rate $\rho_i$ for updating researcher $i$. These $\rho_i$ are collectively subject to the total system update capacity of $c$. We consider the cases of Poisson updates (i.e., updates with exponential inter-update times), deterministic updates, and synchronized updates. We determine the optimal update rates $\rho_i$ subject to the total update rate $c$ in a way to maximize the system freshness. In a broader sense, this problem is related to the problem of real-time timely estimation of signals, which have different change rates and importance factors, with the goal of finding the optimal individual sampling rates, under a total system sampling rate constraint. In this paper, we specialize this broader goal to the setting of counting processes and to the context of tracking citation counts of researchers.

References that are most closely related to our work are [4] and [18]. Reference [4] considers the problem of finding optimal crawl rates to keep the information in a search engine fresh while maintaining the constraints on crawling.

This work was supported by NSF Grants CCF 17-13977 and ECCS 18-07348.
rates imposed by the websites and also the total crawl rate constraint of the search engine, in the presence of non-uniform importance scores and change rates for the websites. Reference [18] focuses on remote real-time reconstruction of a single Poisson counting process using uniform sampling. While taking more samples helps reconstruct the signal better, this increases queuing delays, which inherently affects the real-time signal estimation negatively. Reference [18] studies this trade-off and finds the optimal sampling rate. Their popularity indices. In addition, it was shown in [62] that, the files should be chosen proportional to the square roots of their fresh in a caching system, the (uniform) update rates of files. In particular, it was found in [16] that, in order to keep the returns due to the concavity (sub-linearity) of the square root of the mean citation rate, i.e., there are diminishing optimal update rates are linear in their mean citation rates as researchers are linear in their mean citation rates, then the difference is

$$
\Delta_i(T) = \frac{1}{T} \int_0^T \left( N_i(t) - \hat{N}_i(t) \right) \, dt.
$$

If there are $m$ updates in the interval $[0, T]$, then $\Delta_i(T)$ is

$$
\Delta_i(T) = \frac{1}{T} \sum_{j=1}^m A_j,
$$

where $A_j$ is the difference in the interval $[t_{i,j-1}, t_{i,j}]$, see Fig. 2. Then, the long term average difference for researcher $i$ is $\Delta_i = \lim_{T \to \infty} \Delta_i(T)$, and can be written as [13],

$$
\Delta_i = \lim_{T \to \infty} \frac{1}{T} \sum_{j=1}^m A_j = \lim_{T \to \infty} \frac{m}{T} \cdot \frac{1}{m} \sum_{j=1}^m A_j = \rho_i E[A].
$$

Similar to the derivation in [18], conditioned on an arbitrary jth inter-update time, i.e., $\tau_i = \tau_{i,j} = d$, and the number of citation arrivals in that time interval $[t_{i,j-1}, t_{i,j}+d)$, i.e., $\hat{N}_i(d) = N_i(t_{i,j-1}+d) - N_i(t_{i,j-1}) = k$, the expected difference is

$$
E \left[ A | \hat{N}_i(d) = k, \tau_i = d \right] = kd \frac{d}{2}.
$$

Thus, we have

$$
E \left[ A | \tau_i = d \right] = E \left[ E \left[ A | \hat{N}_i(d) = k, \tau_i = d \right] \right] = \frac{\lambda_i d^2}{2}.
$$

In the following subsections, we present three different models for updating the citation numbers.

**A. Model 1: Poisson Updater**

In this model, shown in Fig. 3, the inter-update times for researcher $i$ are exponential with rate $\rho_i$. Update processes for different researchers are independent. Continuing from (6), we find $E[A]$ using exponential distribution as,

$$
E[A] = \int_0^\infty E[A | \tau_i = t] f_{\tau_i}(t) \, dt = \frac{\lambda_i}{2} E \left[ \tau_i^2 \right] = \frac{\lambda_i}{\rho_i^2}.
$$

Thus, the long term average difference $\Delta_i$ in (4) with a Poisson updater is

$$
\Delta_i = \frac{\lambda_i}{\rho_i}.
$$

Fig. 2 shows $N_i(t)$ and $\hat{N}_i(t)$ with black and blue lines.
B. Model 2: Deterministic Updater

In this model, shown in Fig. 4, the inter-update times are deterministic and chosen optimally. Similar to [16], given that there are \( n \) updates for researcher \( i \) in the time interval \([0, T]\), the optimal inter-update times should be chosen equal to each other, i.e., \( \tau_{i,j} = \frac{T}{n+1} \), for all \( j \). Letting \( T \to \infty \), this update scheme results in uniform sampling with rate \( \rho_i \) for researcher \( i \) where \( \rho_i = \lim_{T \to \infty} \frac{n \tau_i}{T} \). By using \( d_i = \frac{1}{\rho_i} \) and (6), we obtain \( \mathbb{E}[A] = \frac{\lambda_i}{2\rho_i} \). Thus, the long term average difference \( \Delta_i \) in (4) with a deterministic (and uniform) updater is

\[
\Delta_i = \frac{\lambda_i}{2\rho_i}. \tag{9}
\]

C. Model 3: Common Synchronized Probabilistic Updater

In this model, shown in Fig. 5, the updater has a common synchronized update schedule that applies to all researchers. The inter-update times of the common updater are exponential with rate \( \rho \). At each update instant, researcher \( i \) is updated with probability \( p_i \). Thus, inter-update times for researcher \( i \) are exponential with rate \( \rho p_i \). Note that here, we create the Poisson updates for researcher \( i \) by thinning the Poisson common updates using probabilistic updates according to \( p_i \). The main problem, therefore, is to choose \( p_i \) for each researcher as it determines its mean update rate. This problem is the same as the one in Section II-A and in the optimal policy \( p_i = t_i \) assuming \( \rho \) is sufficiently large to have feasible \( p_i \), i.e., \( 0 \leq p_i \leq 1 \), for all \( i \).

D. Problem Formulation

Researcher \( i \) has the mean citation rate \( \lambda_i \). In addition, and optionally, we consider an importance factor, \( \mu_i \), for researcher \( i \). This may be removed by choosing all \( \mu_i = \mu \). Then, the total long term average difference (over all researchers) becomes

\[
\Delta = \sum_{i=1}^{n} \mu_i \Delta_i, \tag{10}
\]

where per researcher difference, \( \Delta_i \), is given by (8) for the Poisson updater and common synchronized updater models, and by (9) for the deterministic updater model. The expressions in (8) and (9) differ only by a factor of 2, which is inconsequential for optimization purposes. Therefore, without loss of generality, from now on, we use the expression in (8). In addition, due to computational limitations, the updater is subject to a total update rate constraint \( \sum_{i=1}^{n} \rho_i \leq c \).

Our aim is to find the optimal update rates for all researchers, \( \rho_i \), for \( i = 1, \ldots, n \), such that the total long term average difference, \( \Delta \), is minimized while satisfying the constraint on the total update rate. Thus, our optimization problem is,

\[
\begin{align*}
\min_{\{\rho_i\}} & \quad \sum_{i=1}^{n} \frac{\mu_i \lambda_i e^{-\rho_i}}{\rho_i} \\
\text{subject to} & \quad \sum_{i=1}^{n} \rho_i \leq c \\
& \quad \rho_i \geq 0, \quad i = 1, \ldots, n. \tag{11}
\end{align*}
\]

We solve the optimization problem in (11) in the next section.

III. THE OPTIMAL SOLUTION

The optimization problem in (10) is convex as the cost function is convex and the constraints are linear. We introduce the Lagrangian function [63] for (10) as

\[
\mathcal{L} = \sum_{i=1}^{n} \frac{\mu_i \lambda_i}{\rho_i} + \beta \left( \sum_{i=1}^{n} \rho_i - c \right) - \sum_{i=1}^{n} \nu_i \rho_i, \tag{12}
\]

where \( \beta \geq 0 \) and \( \nu_i \geq 0 \) for all \( i \). Next, we write the KKT conditions as

\[
\frac{\partial \mathcal{L}}{\partial \rho_i} = -\frac{\mu_i \lambda_i}{\rho_i^2} + \beta - \nu_i = 0, \tag{13}
\]

for all \( i \), and the complementary slackness conditions as

\[
\begin{align*}
\beta \left( \sum_{i=1}^{n} \rho_i - c \right) &= 0, \tag{14}
\nu_i \rho_i &= 0,
\end{align*}
\]

for all \( i \). Since the optimization problem in (10) is convex, the KKT conditions are necessary and sufficient. First, we observe that the total update rate constraint \( \sum_{i=1}^{n} \rho_i \leq c \) must be satisfied with equality. If there is an update rate allocation policy such that \( \sum_{i=1}^{n} \rho_i < c \), then we can achieve a lower average difference by increasing any \( \rho_i \).
as the cost function of (10) is a decreasing function of \( \rho_i \). Thus, in the optimal update rate allocation policy, we must have \( \sum_{i=1}^{n} \rho_i = c \) and \( \beta \geq 0 \) due to (13).

Next, we note that in the optimal policy, we must have \( \rho_i > 0 \), for all \( i \), as \( \rho_i = 0 \) leads to infinite objective function in (10) which clearly cannot be an optimal solution. Thus, for the optimal rate allocations, we have \( \rho_i > 0 \) and \( \nu_i = 0 \), for all \( i \), due to (14).

From (12), we find \( \rho_i = \sqrt{\frac{\mu_i \lambda_i}{c}} \). By using \( \sum_{j=1}^{n} \rho_i = c \), we solve \( \beta = \left( \frac{\sum_{i=1}^{n} \mu_i \lambda_i}{\sum_{i=1}^{n} \sqrt{\mu_i \lambda_i}} \right)^2 \), which gives the optimal policy,

\[
\rho_i = \frac{c \sqrt{\mu_i \lambda_i}}{\sum_{j=1}^{n} \sqrt{\mu_j \lambda_j}}, \quad i = 1, \ldots, n. \tag{15}
\]

Using the optimal rate allocation policy in (15), we obtain

\[
\Delta_i = \frac{\sqrt{\lambda_i} \left( \sum_{j=1}^{n} \sqrt{\mu_j \lambda_j} \right)}{c \sqrt{\mu_i}}, \quad i = 1, \ldots, n, \tag{16}
\]

and the total long term average difference \( \Delta \) as

\[
\Delta = \left( \frac{\sum_{j=1}^{n} \sqrt{\mu_j \lambda_j}}{c} \right)^2. \tag{17}
\]

Thus, the optimal update rates allocated to researchers in (15) are proportional to the square roots of their importance factors, \( \mu_i \), multiplied by their mean citation rates, \( \lambda_i \). We note that if we ignore the importance factors, i.e., \( \mu_i = \mu = 1 \), then the optimal update rates are proportional to the square roots of the mean citation rates. On the other hand, if we choose the importance factors as proportional to the mean citation rates, i.e., \( \mu_i = \alpha \lambda_i \), then the optimal update rates become linear in the mean citation rates.

### IV. Numerical Results

In this section, we provide three numerical results. In the first two examples, we choose the mean citation rates as

\[
\lambda_i = ar^i, \quad i = 1, \ldots, n, \tag{18}
\]

where \( a > 0 \) and \( 0 < r \leq 1 \).

In the first example, we take \( a = 10 \), \( r = 0.75 \), \( n = 20 \) and \( c = 10 \). For this example, we use uniform importance coefficients, i.e., \( \mu_i = \mu = 1 \), for all \( i \). We observe in Fig. 6(b) that researchers with higher mean citation rates have higher long term average difference \( \Delta_i \), even though they are updated with higher update rates shown in Fig. 6(a). Further, we observe in Fig. 6(a) that due to diminishing returns caused by the square root allocation policy, update rates of the researchers with low mean citation rates are still comparable to the update rates of the researchers with high mean citation rates.

In the second example, we consider the case where the importance factors are chosen proportional to the mean citation rates of the researchers. We call such coefficients as \( \lambda \)-proportional importance coefficients, which are given by

\[
\mu_i = \frac{\lambda_i}{\sum_{j=1}^{n} \lambda_j}, \quad i = 1, \ldots, n. \tag{19}
\]

In this third example, we choose the mean citation rates as

\[
\lambda_i = \frac{ar^i}{\sum_{j=1}^{n} r^j}, \quad i = 1, \ldots, n, \tag{20}
\]

which satisfy \( \sum_{i=1}^{n} \lambda_i = a \). Note that, by this selection, we force total citation means of all researchers to be a constant. For this example, we take \( n = 10 \), \( a = 1 \) and consider three different \( r \), which are \( r = 0.5, 0.75, 1 \). Note also that, a smaller \( r \) corresponds to a less even (more polarized) distribution of total mean citation rates among the researchers. We use uniform importance coefficients and plot achieved \( \Delta \) with respect to \( c \) in Fig. 8. We observe in Fig. 8 that more polarized distribution of mean citation rates (smaller \( r \)) yields a lower \( \Delta \) for the system, as we exploit the differences among the researchers by
be withdrawn or changed, which may not be the case, espe-

The monotonically increasing nature of the counting process is monotonically increasing and increments one at a time. citation numbers of a researcher as a counting process, which future research directions. First, we note that we modeled the citation rates multiplied by their importance factors (if any).

The optimal policy is to choose the update rates of individual counts by a resource-constrained updater. We showed that the optimal policy is to choose the update rates of individual counts by a resource-constrained updater. We showed that the achieved importance coefficients are used and \( \lambda_i \) given in (18), with \( a = 10 \) and \( r = 0.75 \) for \( n = 10 \).

allocating even higher update rates to researchers with higher mean citation rates. As an aside, we note that if we used \( \lambda \)-proportional importance coefficients, we would have a \( \Delta \) which is independent of individual \( \lambda_i \) that depends only on the sum of \( \lambda_i \) which is \( a \) here. This achieved \( \Delta \) is also equal to the \( \Delta \) achieved with uniform importance coefficients when \( r = 1 \) which is shown as the blue dashed line in Fig. 8. Thus, if we use \( \lambda \)-proportional importance coefficients, the achieved \( \Delta \) is independent of the mean citation rate distribution among the researchers, but it results in higher \( \Delta \). In other words, uniform importance coefficients achieve lower \( \Delta \) compared to \( \lambda \)-proportional importance coefficients in this case for \( r < 1 \).

V. Conclusion, Discussion and Future Directions

We considered the problem of timely updating of citation counts by a resource-constrained updater. We showed that the optimal policy is to choose the update rates of individual researchers proportional to the square roots of their mean citation rates multiplied by their importance factors (if any).

Next, we discuss limitations of our model and suggest future research directions. First, we note that we modeled the citation numbers of a researcher as a counting process, which is monotonically increasing and increments one at a time. The monotonically increasing nature of the counting process assumes that published articles are always available and cannot be withdrawn or changed, which may not be the case, espe-

\[ \phi \]

\[ \Delta \]}

Fig. 8. Total long term average difference \( \Delta \) with respect to \( c \), when uniform importance coefficients are used and \( \lambda_i \) are given in (20), with \( a = 1 \) and \( r = 0.5, 0.75, 1 \) for \( n = 10 \).

Finally, we assumed that mean citation rates and importance factors are fixed and known by the updater. As a future direction, an online setting can be considered where both of these parameters are learned from observations over time.

REFERENCES


