Computing the Hit Rate of Similarity Caching

Younes Ben Mazziane, Sara Alouf, Giovanni Neglia, Daniel Sadoc Menasche

Computing the Hit Rate of Similarity Caching

Younes Ben Mazziane*, Sara Alouf*, Giovanni Neglia* and Daniel Sadoc Menasche†
*Université Côte d'Azur, Inria, France, email: name.surname@inria.fr
†Federal University of Rio de Janeiro, UFRJ, Brazil, email: sadoc@dcc.ufrj.br

***Abstract*—Similarity caching allows requests for an item i to be served by a similar item i'. Applications include recommendation systems, multimedia retrieval, and machine learning. Recently, many similarity caching policies have been proposed, but still we do not know how to compute the hit rate even for the simplest policies, like SIM-LRU and RND-LRU that are straightforward modifications of classical caching algorithms. This paper proposes the first algorithm to compute the hit rate of similarity caching policies under the independent reference model for the request process. In particular, our work shows how to extend the popular TTL approximation from classic caching to similarity caching. The algorithm is evaluated on both synthetic and real world traces.**

***Index Terms*—Caching, TTL approximation, performance evaluation.**

I. Introduction

Many applications require to retrieve items similar to a given user's request. For example, in content-based image retrieval [1] systems, users can submit an image to obtain other visually similar images. A similarity cache may intercept the user's request, perform a local similarity search over the set of locally stored items, and then, if the search result is evaluated satisfactory, provide it to the user. The cache thus may speed up the reply and reduce the load on the server, at the cost of providing items *less similar* than those the server would provide. Originally proposed for content-based image retrieval [1] and contextual advertising [2], similarity caches are now a building block for a large variety machine learning based inference systems for recommendations [3], image recognition [4], [5] and network traffic [6] classification. In these cases, the similarity cache stores past queries and the respective inference results to serve future similar requests.

Motivated by the large number of applications, recently much effort has been devoted to formalize similarity caching [7], [8] as well as to propose new caching policies [9]–[11]. Despite this research, to the best of our knowledge, we still do not know how to compute basic performance metrics—like the percentage of requests satisfied by the cache—even for the simplest similarity caching policies, like SIM-LRU and RND-LRU, which were proposed in the seminal paper [2]. SIM-LRU and RND-LRU are variants of the basic LRU policy, but are much more challenging to analyze than LRU due to strong coupling across items in the cache. In fact, in classic caching, an item in the cache only contributes to serve requests for the very same item, while in similarity caching, the same item can serve requests for a set of similar items as far as neither them, nor their most similar items, are stored in the cache. It follows that, in similarity caching, the number of requests satisfied by an item in the cache depends in general on *the whole cache state*.

TABLE I: Table of notation

Variable	Description
Basic parameters	
I	set of items
$N = \|I\|$	catalogue size
C	cache capacity
S	state of cache; set of cached items
λ_n	arrival rate of requests to item n
Similarity cache parameters	
$\mathrm{dis}(\cdot,\cdot)$	function measuring the dissimilarity between items
d	threshold similarity
$\mathcal{N}(n)$	neighbours of item n
$\mathcal{N}[n]$	neighbours of item n including n
$\mathcal{N}_i(n)$	items in $\mathcal{N}(n)$ strictly closer to n than i
$\mathcal{N}_i[n]$	items in $\mathcal{N}[n]$ strictly closer to n than i
$q_n(i)$	probability of an approximate hit for i by n
Inferred variables	
λ_n^e	insertion rate of item n; rate at which it enters the cache
λ_n^r	refresh rate of item n
t_C	characteristic time
Key metrics of interest	
h_n	probability of an approximate hit of a request for item n
o_n	fraction of time that item n is cached
H	cache hit probability

In this paper, we introduce the first algorithm to estimate SIM-LRU and RND-LRU hit rate under the independent reference model (IRM) [12] for the request process. The algorithm alternates between two steps. In the first step, given a tentative estimate of the rate of requests served by each item when present in the cache, the occupancy probability of each item (i.e., the probability that the item is in the cache) is computed relying on the well-known TTL approximation [12], [13] (also known as Che's approximation), which has been successfully used to study classic caching policies. In the second step, the current vector of occupancy probabilities, and similarity relations across items, are used to update the rate of requests served by each item. Our experiments both on synthetic traces and a realistic trace for a recommendation system show that our algorithm provide accurate estimates of the hit rate, definitely more precise than other intuitive approaches one could think about.

The paper is organized as follows: background and notation are introduced in Sec. II, our algorithm for computing the hit rate for SIM-LRU and RND-LRU is presented in Sec. III, its performance is evaluated on both synthetic and real word traces in Sec. IV and Sec. V concludes.

II. Background, Notation and Assumptions

A. Similarity Caching

In **similarity search** systems users can request to a remote server, storing a set of items I, the k most similar items to a given item n, given a specific definition of similarity. In practice items are often represented by vectors in $\mathbb{R}^d$ (called embeddings) [14] so that the dissimilarity cost, $\mathrm{dis}(.,.) : I^2 \to \mathbb{R}^+$, can be selected to be an opportune distance between the embeddings, e.g., the Euclidean one. A cache, that stores a small fraction of the catalog I, could be deployed next to the users to reduce the fetching cost of similarity searches. The seminal papers [1], [2] suggest the cache may answer a request using a local subset of items potentially different from the true closest neighbors to further reduce the fetching cost while still maintaining an acceptable dissimilarity cost. They refer to such caches as **similarity caches**.

One of the popular dynamic similarity caching policies is SIM-LRU [2]. This policy maintains an ordered list of C key-value pairs. Each key is the embedding of an item n requested in the past and its corresponding value is a list containing the $k' \geq k$ closest items to n in I. We denote by S the set of keys stored in the cache. Upon a similarity search for an item n, SIM-LRU selects the closest local key to n, i.e., $\hat{n} \triangleq \arg\min_{m \in S} \mathrm{dis}(n, m)$. If the dissimilarity cost between n and $\hat{n}$ is smaller than a threshold $d > 0$ ($\mathrm{dis}(n, \hat{n}) \leq d$), the request experiences an *approximate hit*:[1] the cache replies to the request for n selecting the k closest items to n among the k' values stored for $\hat{n}$ and moves $\hat{n}$'s key-value pair to the front of the list. Otherwise, the request experiences a *miss*: it is forwarded to the original server to retrieve the k' closest items to n, out of which the closest k are provided to the user. The cache then adds the new key-value pair for n to the front of the list and evicts the key-value pair at the bottom of the list. We observe how the use of key-value pairs in SIM-LRU essentially converts the search of k closest items into the search of the closest key in the cache. For simplicity's sake, from now on we will just identify the items, their keys and the corresponding values and say for example that the cache replies to a request for n with the closest item $\hat{n}$ in the cache.

RND-LRU [2] is a generalization of SIM-LRU, where $\hat{n}$ is used to reply to a query for n with a probability $q_{\hat{n}}(n)$ which decreases with their dissimilarity, and it is in any case null for $\mathrm{dis}(\hat{n}, n) > d$. We retrieve the behaviour of SIM-LRU when $q_{\hat{n}}(n) = 1$ if $\mathrm{dis}(\hat{n}, n) \leq d$.

B. Our Assumptions

We assume that requests follow a Poisson process with request rate λ_n for item n, and each request is independent from the previous ones, i.e., requests follow the Independent Reference Model (IRM) [12]. Under SIM-LRU or RND-LRU, a request for item n could be served by any item closer than d to n. We denote the set of such items as $\mathcal{N}[n] \triangleq \{m \in I : \ \mathrm{dis}(n, m) \leq d\}$. We call the elements in $\mathcal{N}[n]$ distinct from n the *neighbours* of n and we denote their set as $\mathcal{N}(n) \triangleq \mathcal{N}[n] \setminus \{n\}$. For the sake of simplicity, we assume that items in $\mathcal{N}(n)$ can be strictly ordered according to their dissimilarity wrt n, i.e., for any $(i, j) \in \mathcal{N}(n)$ and $i \neq j$, we have $\mathrm{dis}(n, i) \neq \mathrm{dis}(n, j)$. If this is not the case, we can introduce an arbitrarily order for items with the same dissimilarity. For convenience, we also define in a similar way the sets $\mathcal{N}_i(n)$ and $\mathcal{N}_i[n]$, subsets of $\mathcal{N}(n)$ and $\mathcal{N}[n]$, resp., designating items that are closer to n than i, i.e. $\mathcal{N}_i(n) \triangleq \{m \in \mathcal{N}(n) : \ \mathrm{dis}(n, m) < \mathrm{dis}(n, i)\}$ and $\mathcal{N}_i[n] \triangleq \{m \in \mathcal{N}[n] : \ \mathrm{dis}(n, m) < \mathrm{dis}(n, i)\}$.

[1]Note that we have an exact hit if $\hat{n} = n$.

C. TTL Approximation for LRU Cache

The hit rate of an LRU cache under the IRM model can be estimated using what is referred to in the literature as the TTL approximation [12], [13]. The approximation considers that any cached item n, if not requested, will stay in an LRU cache with capacity C for a time t_C that is deterministic and independent of n; t_C is called the cache 'characteristic time'. This approximation has been later supported by theoretical arguments in [15], [16]. Under the TTL approximation, a hit occurs for an item if the inter arrival time between two requests for the same item is smaller than t_C. Thus, the hit probability h_n can be approximated as:

$$h_n \approx 1 - \mathrm{e}^{-\lambda_n t_C} \ . \tag{1}$$

Since the flow of arrivals is Poisson, the "Poisson Arrivals See Time Averages" (PASTA) property implies that the probability o_n that an item n is in the cache (i.e., the occupancy probability, or simply occupancy) is equal to the probability that a request for that same item experiences a hit, i.e. $h_n = o_n$. The cache capacity constraint is given in expectation by the following equality:

$$\sum_{n \in I} o_n = C \ , \tag{2}$$

The above expression allows us to deduce t_C, e.g., by means of a bisection method. The hit rate H can be simply computed as $H = \sum_n \lambda_n h_n$ with h_n computed as in (1).

III. TTL Approximation for Similarity Caching

Analogously to LRU, TTL approximation for RND-LRU assumes that the time an item stays in the cache if it is not serving any requests is deterministic and independent of n and we denote it as t_C. The hit rate for an item n (i.e., the rate of requests incurring an approximate hit) can no longer be computed as in (1) as the request for n can be satisfied even if n is not in the cache. Let S denote the current state of the cache, i.e., the set of items it stores. For RND-LRU, an approximate hit for item n occurs if at least one of the items in $\mathcal{N}[n]$ is present in the cache. When the closest item to n present in the cache is $i \in \mathcal{N}[n]$, i.e. $S \cap \mathcal{N}_i(n) = \emptyset$ and $i \in S$, i serves the request for n with probability $q_i(n)$. Taking advantage of the PASTA property, it follows that h_n for RND-LRU can be expressed as:

$$h_n = \sum_{i \in \mathcal{N}[n]} q_i(n) \cdot \Pr\left(S \cap \mathcal{N}_i(n) = \emptyset, i \in S\right). \tag{3}$$

In what follows, we provide an alternative approach to compute the occupancies for an LRU cache, under the TTL approximation, that is complementary to that presented in Sec. II-C. Then, we identify the differences between RND-LRU and LRU and compute its occupancies in a similar way.

A. Occupancies and Hit Rates

To derive the occupancy of an item n, we first observe that the instants when item n is evicted from the cache are regeneration points for a renewal process [17]. A renewal cycle consists of two consecutive time periods: a time period of duration T_n^{Off}, that starts immediately after item n is evicted from the cache and ends when it re-enters the cache, and a time period of duration T_n^{On}, that ends when item n is evicted again from the cache. From the renewal theorem, the occupancy can be computed as:

$$o_n = \frac{\mathbb{E}\left[T_n^{\text{On}}\right]}{\mathbb{E}\left[T_n^{\text{Off}}\right] + \mathbb{E}\left[T_n^{\text{On}}\right]} \ . \tag{4}$$

a) Expectation of T_n^{Off}: T_n^{Off} is the waiting time for a miss for n after n has been evicted from the cache. For LRU, under the IRM model, thanks to the memoryless property of the exponential distribution the residual interarrival time T_n^{Off} is exponentially distributed with rate λ_n, implying that $\mathbb{E}\left[T_n^{\text{Off}}\right] = 1/\lambda_n$.

For RND-LRU, when a request for n finds none of item n's neighbours in the cache, a miss occurs with probability 1. If instead a request for n finds i to be the closest neighbour of n in the cache, i.e. $\mathcal{N}_i(n) \cap S = \emptyset, i \in S$, the probability of a miss is $1 - q_i(n)$. Let $p_n^e(i)$ be the probability that i is the closest neighbour to n in the cache, and that a miss occurs, namely:

$$p_n^e(i) \triangleq (1 - q_n(i)) \Pr\left(S \cap \mathcal{N}_i(n) = \emptyset, i \in S \mid n \notin S\right). \tag{5}$$

The probability p_n^e of a miss for n is then:

$$p_n^e \triangleq \Pr\left(S \cap \mathcal{N}(n) = \emptyset \mid n \notin S\right) + \sum_{i \in \mathcal{N}(n)} p_n^e(i) \ . \tag{6}$$

Consequently, when item n is not cached, the rate at which item n re-enters the cache is:

$$\lambda_n^e = \lambda_n p_n^e \ . \tag{7}$$

Similarly to the case of LRU, we write $\mathbb{E}\left[T_n^{\text{Off}}\right] = 1/\lambda_n^e$.

b) Expectation of T_n^{On}: For LRU, the time spent by an item in the cache is at least t_C, under TTL approximation. Each time the interarrival time between requests for item n is smaller than t_C, there is a hit, and n is moved to the top of the queue. In this case, item n is refreshed, i.e., its timer is re-initialized. On the other hand, when this interarrival time is larger than t_C, n is evicted from the cache. Time interval T_n^{On} is the sum of a random number F of time intervals with duration shorter than t_C (corresponding to F consecutive hits) and t_C (the time between the last hit and item's eviction). It follows that

$$T_n^{\text{On}} \approx \sum_{j \in [1..F]} X_j + t_C \ , \tag{8}$$

where $(X_j)_{j \in [1..F]}$ are the interarrival times between requests for item n such that $X_j < t_C$ for all $j \in [1..F]$. Since we have:

$$\mathbb{E}\left[X_j \mid X_j < t_C\right] = \frac{1}{\lambda_n} - \frac{t_C}{\exp(\lambda_n t_C) - 1} \ , \tag{9}$$

$$\mathbb{E}\left[F\right] = \exp(\lambda_n t_C) - 1 \ , \tag{10}$$

we conclude from Wald's identity and (8) that:

$$\mathbb{E}\left[T_n^{\text{On}}\right] \approx \frac{\mathrm{e}^{\lambda_n t_C} - 1}{\lambda_n} \ . \tag{11}$$

For RND-LRU, item n is not only refreshed by its own requests but also by requests for its neighbours. More specifically, if item $i \in \mathcal{N}[n]$ is requested and item n is the closest neighbour to i among all cached items, then item n first 1) serves the request for i with probability $q_n(i)$ and then 2) is refreshed in the cache. In such a case, the state of the cache verifies $S \cap \mathcal{N}_n[i] = \emptyset$. T_n^{On} can still be written as in (8). However, the refresh rate for random variables X_j is no longer λ_n but a higher rate λ_n^r expressed as:

$$\lambda_n^r = \sum_{i \in \mathcal{N}[n]} p_n^r(i) \lambda_i, \tag{12}$$

$$p_n^r(i) = q_n(i) \cdot \Pr\left(S \cap \mathcal{N}_n[i] = \emptyset \mid n \in S\right) \ . \tag{13}$$

(Notice that $\mathcal{N}_n[n] = \emptyset$ and then $p_n^r(n) = 1$.) Similarly to (11), the expected value of T_n^{On} can be computed as

$$\mathbb{E}\left[T_n^{\text{On}}\right] \approx \frac{\mathrm{e}^{\lambda_n^r t_C} - 1}{\lambda_n^r} \ . \tag{14}$$

c) Computing the occupancy: Replacing the expressions for $\mathbb{E}\left[T_n^{\text{Off}}\right]$ and $\mathbb{E}\left[T_n^{\text{On}}\right]$ in (4), we derive the occupancy. For LRU, we obtain the knonw result in (1) (remember that $h_n = o_n$ for LRU). For RND-LRU, we obtain

$$o_n \approx \left(\frac{1}{\lambda_n^e} \cdot \frac{\lambda_n^r}{\mathrm{e}^{\lambda_n^r t_C} - 1} + 1\right)^{-1} . \tag{15}$$

d) Computing p_n^e, $p_n^r(i)$ and the hit rate: Under the classic TTL approximation for LRU, items are coupled only through the value of the characteristic time t_C. Conditioned on t_C, events related to the presence of items in the cache are independent, e.g., $\Pr\left(n, m \in S\right) = \Pr\left(n \in S\right) \cdot \Pr\left(m \in S\right)$. We maintain this independence also for RND-LRU. Then, (6), (5), and (13) can be written as follows:

$$p_n^e(i) = (1 - q_i(n)) \cdot o_i \prod_{m \in \mathcal{N}_i(n)} (1 - o_m) \ , \tag{16}$$

$$p_n^e = \prod_{m \in \mathcal{N}(n)} (1 - o_m) + \sum_{i \in \mathcal{N}(n)} p_n^e(i) \ , \tag{17}$$

$$p_n^r(i) = q_n(i) \prod_{m \in \mathcal{N}_n[i]} (1 - o_m) \ . \tag{18}$$

Under the independence assumption, RND-LRU's hit rate h_n for item n's requests in (3) can be computed as:

$$h_n = \sum_{i \in \mathcal{N}[n]} q_i(n) \cdot o_i \prod_{m \in \mathcal{N}_i(n)} (1 - o_m) \ . \tag{19}$$

Algorithm 1: Fixed point method

Input: C, $\vec{\lambda}$, dis(.,.), d, $(q_n(i))_{(n,i)\in I^2}$, stopping condition
Output: Estimation $\vec{o}, \vec{h}, t_C$
Initialization:
1: Obtain $t_C(0)$ such that $\sum_{n\in I}\left(1-\mathrm{e}^{-\lambda_n \cdot t_C(0)}\right) = C$
2: $\vec{o}(0) \leftarrow 1-\mathrm{e}^{-\vec{\lambda}\cdot t_C(0)}$
3: $\vec{h}(0) \leftarrow f^h(\vec{o}(0))$
4: $j \leftarrow 1$
5: **while** *Stopping condition not satisfied* **do**
6: $\quad \vec{\lambda^e}(j) \leftarrow f^e(\vec{o}(j-1))$ (See (20))
7: $\quad \vec{\lambda^r}(j) \leftarrow f^r(\vec{o}(j-1))$ (See (21))
8: $\quad$ Obtain $t_C(j)$ such that : $\sum_{n\in I}(f^o(\vec{\lambda^e}(j), \vec{\lambda^r}(j), t_C(j)))_n = C$ (See (23),(22))
9: $\quad \vec{o}(j) \leftarrow (f^o(\vec{o}(j-1), t_C(j)) + \vec{o}(j-1))/2$
10: $\quad \vec{h}(j) = f^h(\vec{o}(j))$ (See (24))
11: $\quad j \leftarrow j+1$
12: **end while**
13: **return** $\vec{h}(j)$, $\vec{o}(j)$, $t_C(j)$

B. Algorithm for Finding Hit Probabilities

Next, our goal is to propose an algorithm to compute hit probabilities. To this aim, we solve the following set of equations:

$$\vec{\lambda^e} = f^e(\vec{o}) \ , \tag{20}$$

$$\vec{\lambda^r} = f^r(\vec{o}) \ , \tag{21}$$

$$\vec{o} = f^o(\vec{\lambda^r}, \vec{\lambda^e}, t_C), \tag{22}$$

$$\sum_{n\in I} o_n = C, \tag{23}$$

$$\vec{h} = f^h(\vec{o}) \ . \tag{24}$$

Equation (20) follows from (7), (16), and (17), and computes the vector of insertion rates for all items. Equation (21) follows from (12) and (18) and computes the vector of refresh rates. Equation (22) is the vector form of (15): given $\vec{\lambda^r}$ and $\vec{\lambda^e}$, and the characteristic time t_C, it computes all occupancies. Equation (23) expresses the capacity constraint. Combining (20)–(23), we obtain a system of $3N+1$ equations in $3N+1$ unknowns, from which we can obtain in particular the occupancies and the characteristic time t_C. Finally, once the occupancies are known Equation (24) computes the vector of hit rates according to (19).

To solve the system of equations (20)-(23), we rely on an iterative fixed point method (see Algorithm 1). We begin by guessing occupancies $\vec{o}$. In particular, we initialize them using occupancies for LRU, i.e., $\vec{o}(0) = 1-\mathrm{e}^{-\vec{\lambda}\cdot t_C(0)}$ where $t_C(0)$ verifies (2) ($\sum_{n\in I} o_n(0) = C$) (lines 1-2). Then, we obtain $\vec{\lambda^e}(1)$ and $\vec{\lambda^r}(1)$ using equations (20) and (21), resp. (lines 5-7). Next we find the new estimation of the occupancies $\vec{o}(1) = f^o(\vec{\lambda^e}(j), \vec{\lambda^r}(j), t_C(1))$ where $t_C(1)$ verifies $\sum_{n\in I} o_n(1) = C$ (lines 8-9). Finally, a new estimate of the vector of occupancies is computed (line 9): averaging the new prediction and the previous value is a practical trick to improve the convergence. The same procedure is then repeated for the next iterations until a stopping condition is reached, e.g., the difference between $\vec{o}$ computed at consecutive iterations becomes smaller than a given threshold, or the maximum number of iterations is reached ($j \leq n_{\text{iterations}}$).

C. Benchmarks and Alternative Approaches

In what follows, we compare hit rate estimates provided by RND-LRU or SIM-LRU using Algorithm 1 with the hit rate estimations for LRU and for the optimal static allocation. We also propose an alternative approach to estimate RND-LRU's hit rate.

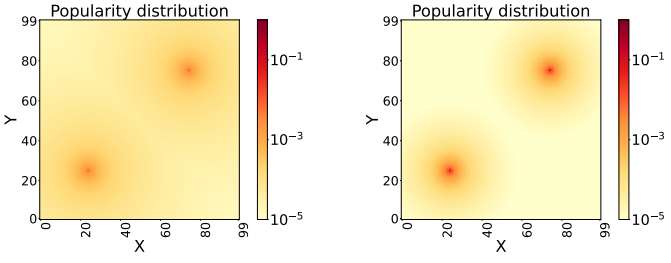
LRU. The hit rate and the occupancy for an item n are computed using Eq. (1) and t_C is deduced using the cache capacity constraint given by Eq. (2).

Optimal Static Allocation. The maximum hit rate obtainable by a static allocation under similarity caching can be obtained solving a maximum weighted coverage problem. We consider, as in SIM-LRU, that each item can be used to satisfy any request for items closer than d. The maximum weighted coverage problem takes as input a capacity C, a set of items I, with $N = |I|$, their corresponding weights $W = (w_i)_{i\in I}$ and a set of sets $R = \{R_1, \ldots, R_N\}$ such that $R_i \subset I$. The objective is to find a set $\sigma^* \subset \{1, \ldots, N\}$ such that: $\sigma^* = \arg\max_{\sigma\subset\{1,\ldots,N\}:|\sigma|\leq C} \sum_{i\in\cup_{j\in\sigma} R_j} w_i$. Finding the best static allocation is equivalent to solving a maximum weighted coverage problem, with weights $w_i = \lambda_i$ for $i \in I$, C the cache capacity, and R the set of neighbours for each item, i.e., $R = \{\mathcal{N}[n]\}_{n\in I}$. The maximum weighted coverage problem is known to be NP-hard. In practice, a popular greedy algorithm guarantees a $(1-1/e)$ approximation ratio [18].

The greedy algorithm chooses at the first step the set with the largest coverage $c_m = \max_{n\in I} \sum_{i\in R_n^0 = R_n} p_i$. If R_o^0 is the set chosen at the first iteration, at the next iteration all the sets are updated in such a away that they do not contain any item in the set R_o^0, i.e. $R_n^1 = R_n^0 \setminus R_o^0$. The same procedure is repeated until C sets are collected or all the sets are chosen.

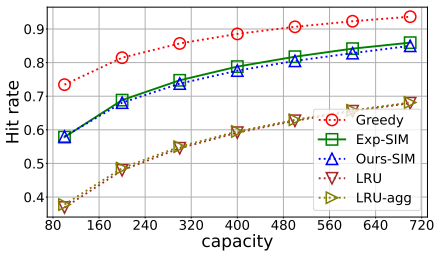
LRU with aggregate requests. Under SIM-LRU an item is refreshed by the requests for all its neighbours. A naive approach to study a SIM-LRU cache is then to consider that it operates as a LRU cache with equivalent request rates for each item equal to the sum of the request rates for all items in its neighborhood. One can then use the TTL approximation for LRU, leading to the following formulas:

$$h_n = 1-\mathrm{e}^{-\sum_{i\in\mathcal{N}[n]}\lambda_i\cdot t_C}, \quad o_n = h_n \ . \tag{25}$$

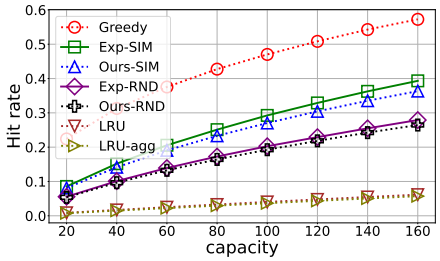


(a) $\alpha = 1.4$ (b) $\alpha = 2.5$

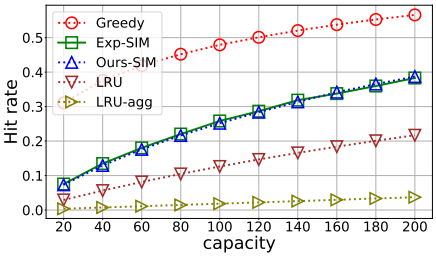
Fig. 1: Synthetic traces: Spatial popularity distribution.



(a) Synthetic trace, $\alpha = 2.5$, $d = 1$, 25 iterations



(b) Synthetic trace, $\alpha = 1.4$, $d = 2$, 15 iterations



(c) Amazon trace, $d = 300$, 40 iterations

Fig. 2: Hit rate versus cache capacity.

IV. Numerical Evaluation

We evaluate the efficiency of the proposed fix point method (Algorithm 1) to predict the hit rate on synthetic traces and on an Amazon trace [10]. For the synthetic traces, each item corresponds to two features, characterized by a point in a grid, $I = [0..99]^2$ (e.g. Fig. 1). The total number of items is $|I| = 10^4$, and the dissimilarity function between items $\mathrm{dis}(\cdot,\cdot)$ is the Euclidean distance. Neighbours of item (x, y) at the same distance are ordered counterclockwise starting from the item to the right, i.e., from $(x + a, y)$ with $a > 0$. Note that for similarity thresholds $d \in \{1, 2\}$ the proposed distance produces an ordering equivalent to Manhattan distance (MD), with MD ties broken in such a way that items in same row or column have higher distance than their counterparts.

We generate a stream of r requests for items in I in an IRM fashion [12], $r = 2 \cdot 10^5$. The popularity distribution for an item $n = (x, y)$ is given by

$$p_{(x,y)} \sim \left(\min\left\{\mathrm{dis}(n, (24, 24)), \mathrm{dis}(n, (74, 74))\right\} + 1\right)^{-\alpha}, \tag{26}$$

where α is a parameter controlling the skew of the popularity distribution. Fig. 1 illustrates the cases $\alpha \in \{1.4, 2.5\}$.

For the Amazon trace, [14] proposes a scheme to embed the images of Amazon products in a 100-dimensional space, where the Euclidean distance reflects dissimilarity between two items. Then, [10] reports the number of reviews per product, and equates it to product request rates. Inspired by this methodology, we leverage the empirical request probabilities, and use it to generate a corresponding IRM stream of requests.

Given the workloads, we evaluate similarity cache mechanisms employing SIM-LRU with threshold similarity $d \in \{1, 2\}$ for the synthetic traces and $d = 300$ for the Amazon trace. For the synthetic trace with $d = 2$, we also evaluate RND-LRU where the probabilities $q_n(i)$ are mapped to $\mathrm{dis}(n, i)$ as $\left(\left[1, \frac{1}{2}, \frac{1}{4}\right], \left[1, \sqrt{2}, 2\right]\right)$. The 95% confidence intervals were smaller than $1.2 \cdot 10^{-3}$ in all the considered synthetic experiments for the hit rate computation. In all experiments, we refer to the empirical hit rates for SIM-LRU and RND-LRU as 'Exp-SIM' and 'Exp-RND', respectively.

For all the theoretical computations of the hit rate, the arrival rates $\vec{\lambda}$ for items are taken equal to the corresponding request probabilities. Our approach uses Algorithm 1 to compute the hit rates for each item, $\vec{h}$, and then deduces the cache hit rate H. We refer to the latter estimate, for SIM-LRU and RND-LRU, as 'Ours-SIM' and 'Ours-RND', respectively. Alternative methods that could be used to estimate the hit rate are presented in Sec. III-C. We refer to the TTL approximation for LRU as 'LRU', LRU with aggregate requests as 'LRU-agg', and the greedy algorithm as 'Greedy'. The numerical values used for all the experiments are summarized in Table II.

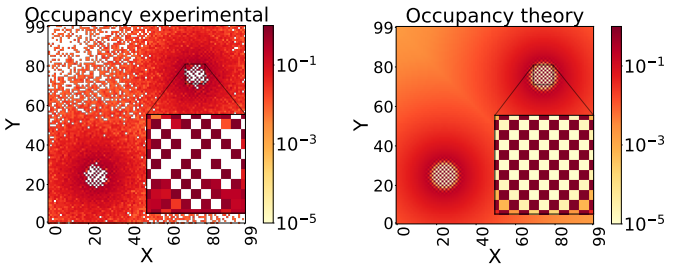
TABLE II: Parameters of the experiments

Variable	Synthetic traces	Amazon trace
I	$[0..99]^2$	Products
$N = \lvert I\rvert$	10^4	$\approx 10^4$
λ_n	(26)	Empirical
$\mathrm{dis}(\cdot,\cdot)$	Euclidean distance	Euclidean distance
d	1 and 2	300
Number of requests r	$2 \cdot 10^5$	$\approx 10^5$
95% confidence intervals	$\approx 10^{-3}$	—
Number of iterations	25 and 15	40

In Fig. 2, we show the empirical hit rate along with its predictions, including those predictions obtained with our approach, for the two synthetic settings and for the Amazon trace. In the considered settings, 'Greedy' overestimates the hit rate. 'LRU' and 'LRU-agg', in contrast, underestimate it.

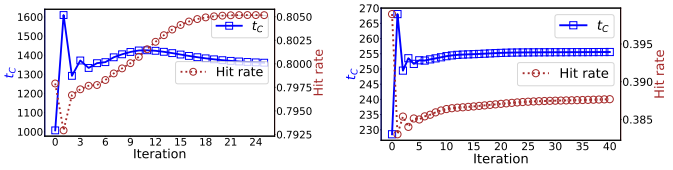
'Ours-SIM' and 'Ours-RND' clearly outperform all the alternative approaches presented in Sec. III-C in estimating the empirical hit rate, while tending to underestimate it. 'LRU' does not take into account the similarity between items, hence the gap between 'LRU' and 'Exp-SIM' shows us the benefits of similarity caching over exact caching. For the synthetic settings in Figs. 2a and 2b, 'LRU' and 'LRU-agg' achieve similar hit rates. This is possibly due to the choice of the popularity distribution (see (26)) where a popular item n and its neighbours have similar popularities: $\vec{\lambda}_n = \sum_{i \in \mathcal{N}[n]} \lambda_i \approx |\mathcal{N}[n]|\lambda_n$, implying that $\vec{\lambda} \approx f(d)\vec{\lambda}$, which corresponds to the case wherein it is equivalent to computate h_n using either 'LRU' or 'LRU-agg'.

To shed further insight on why our approach underestimates the hit rate, Fig. 3 shows the empirically estimated occupancy vector and the one produced by Algorithm 1. The proposed algorithm broadly captures the empirical occupancy patterns, but with subtleties regarding symmetries. In particular, the



(a) $r = 2 \cdot 10^5$ (b) 25 iterations

Fig. 3: Synthetic trace occupancies: $C = 500$, $d = 1$, $\alpha = 2.5$.



(a) Synthetic trace, $\alpha = 2.5$, $d = 1$, $C = 500$ (b) Amazon trace, $d = 300$, $C = 200$

Fig. 4: Characteristic time t_C and hit rate in different iterations of Algorithm 1 for SIM-LRU.

zoom on Fig. 3b shows that our approach produces a regular chess board pattern. Some items are predicted to stay almost all the time in the cache while their 4 neighbours are predicted to spend virtually no time in it. The corresponding empirical occupancy on Fig. 3a shows a less symmetric pattern, implying that in this setup SIM-LRU is able to satisfy a group of requests using a smaller number of cache slots when compared against what is predicted by our approach. This, in turn, partially explains why our approach underestimates the hit rate.

Fig. 4 shows the evolution of characteristic time t_C and hit rate H over different iterations. We observe that estimates of H and t_C by our algorithm converge in few iterations (less than 50), under all the considered scenarios. Note that $t_C(0)$, the value of t_C at iteration 0, is also the value of t_C for 'LRU' (see Eqs. (1) and (2)). In addition, across all experiments, t_C for SIM-LRU using Algorithm 1 converges to a value larger than $t_C(0)$. Indeed, under 'LRU', t_C is bounded by the time required for C distinct items to be requested. For SIM-LRU and RND-LRU, in contrast, after C distinct items are requested, an item previously in cache can remain there, despite not serving any requests. This occurs due to approximate hits, explaining why t_C is larger for SIM-LRU than 'LRU'.

V. Conclusion

We proposed the first algorithm to estimate the hit rate for popular and simple dynamic policies for similarity caching: SIM-LRU and RND-LRU, under the IRM model. Our experimental benchmark shows that our approach outperforms simple methods one can think of to predict the hit rate. Our approach builds on solving a system of equations using a fixed point method. Although our algorithm converged in our experiments, the study of the conditions for convergence is deferred for future work. In addition, note that when using SIM-LRU or RND-LRU two items whose dissimilarity is smaller than d can not be simultaneously cached. We envision to modify our algorithm to take this fact into account. Furthermore, we aim to investigate the asymptotics of the TTL approximation error, similarly to what was done in [15], [16] for classical caches.

Acknowledgement

This project was financed in part by CAPES, CNPq and FAPERJ Grant JCNE/E-26/203.215/2017.

References

[1] F. Falchi, C. Lucchese, S. Orlando, R. Perego, and F. Rabitti, "A Metric Cache for Similarity Search," in *Proc. of 2008 ACM workshop on Large-Scale distributed systems for information retrieval*, pp. 43–50, 2008.

[2] S. Pandey, A. Broder, F. Chierichetti, V. Josifovski, R. Kumar, and S. Vassilvitskii, "Nearest-neighbor caching for content-match applications," in *Proceedings of the 18th international conference on World wide web*, pp. 441–450, 2009.

[3] P. Sermpezis, T. Giannakas, T. Spyropoulos, and L. Vigneri, "Soft cache hits: Improving performance through recommendation and delivery of related content," *IEEE Journal on Selected Areas in Communications*, vol. 36, pp. 1300–1313, June 2018.

[4] U. Drolia, K. Guo, and P. Narasimhan, "Precog: Prefetching for image recognition applications at the edge," in *Proc. of ACM/IEEE Symposium on Edge Computing*, pp. 1–13, 2017.

[5] S. Venugopal, M. Gazzetti, Y. Gkoufas, and K. Katrinis, "Shadow puppets: Cloud-level accurate AI inference at the speed and economy of edge," in *USENIX HotEdge* , 2018.

[6] A. Finamore, J. Roberts, M. Gallo, and D. Rossi, "Accelerating deep learning classification with error-controlled approximate-key caching," in *IEEE Conference on Computer Communications (INFOCOM)*, 2022.

[7] G. Neglia, M. Garetto, and E. Leonardi, "Similarity Caching: Theory and Algorithms," *IEEE/ACM Transactions on Networking*, 2021.

[8] M. Garetto, E. Leonardi, and G. Neglia, "Content placement in networks of similarity caches," *Computer Networks*, vol. 201, p. 108570, 2021.

[9] J. Zhou, O. Simeone, X. Zhang, and W. Wang, "Adaptive offline and online similarity-based caching," *IEEE Networking Letters*, vol. 2, no. 4, pp. 175–179, 2020.

[10] A. Sabnis, T. S. Salem, G. Neglia, M. Garetto, E. Leonardi, and R. K. Sitaraman, "Grades: Gradient descent for similarity caching," in *IEEE Conference on Computer Communications (INFOCOM)*, IEEE, 2021.

[11] T. S. Salem, G. Neglia, and D. Carra, "AÇAI: Ascent Similarity Caching with Approximate Indexes," in *2021 33th International Teletraffic Congress (ITC-33)*, pp. 1–9, IEEE, 2021.

[12] R. Fagin, "Asymptotic miss ratios over independent references," *Journal of Computer and System Sciences*, vol. 14, no. 2, pp. 222–250, 1977.

[13] H. Che, Y. Tung, and Z. Wang, "Hierarchical web caching systems: Modeling, design and experimental results," *IEEE journal on Selected Areas in Communications*, vol. 20, no. 7, pp. 1305–1314, 2002.

[14] J. McAuley, C. Targett, Q. Shi, and A. Van Den Hengel, "Image-based recommendations on styles and substitutes," in *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pp. 43–52, 2015.

[15] C. Fricker, P. Robert, and J. Roberts, "A versatile and accurate approximation for LRU cache performance," in *2012 24th International Teletraffic Congress (ITC 24)*, pp. 1–8, IEEE, 2012.

[16] B. Jiang, P. Nain, and D. Towsley, "On the Convergence of the TTL Approximation for an LRU Cache Under Independent Stationary Request Processes," *ACM Trans. Model. Perform. Eval. Comput. Syst.*, vol. 3, pp. 20:1–20:31, Sept. 2018.

[17] S. I. Resnick, *Heavy-tail phenomena: probabilistic and statistical modeling*. Springer Science & Business Media, 2007.

[18] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular set functions—I," *Mathematical programming*, vol. 14, no. 1, pp. 265–294, 1978.