Information Processing Letters 129 (2018) 35–39

Contents lists available at ScienceDirect

Information Processing Letters

www.elsevier.com/locate/ipl

On the analysis of Bloom filters

Fabio Grandi

Department of Computer Science and Engineering (DISI), Alma Mater Studiorum – Università di Bologna, Viale Risorgimento 2, I-40136 Bologna, BO, Italy

ARTICLE INFO

Article history:
Received 8 August 2017
Accepted 14 September 2017
Available online 21 September 2017
Communicated by Marcin Pilipczuk

Keywords:
Data structures
Analysis of algorithms
Bloom filters
γ-Transform

ABSTRACT

The Bloom filter is a simple random binary data structure which can be efficiently used for approximate set membership testing. When testing for membership of an object, the Bloom filter may give a false positive, whose probability is the main performance figure of the structure. We complete and extend the analysis of the Bloom filter available in the literature by means of the γ-*transform* approach. Known results are confirmed and new results are provided, including the variance of the number of bits set to 1 in the filter. We consider the choice of bits to be set to 1 when an object is inserted both with and without replacement, in what we call *standard* and *classic* Bloom filter, respectively. Simple iterative schemes for the computation of the false positive probability and a new non-iterative approximation, taking into account the variance of bits set to 1, are also provided.

1. Introduction

The Bloom filter [1] is a simple random data structure which can be efficiently used for approximate set membership testing. Considering n objects $o_i \in O$ $(i \in \{1..n\})$ to be inserted in a Bloom filter made of m bits initially set to 0, k independent hash functions $h_j : O \to \{1..m\}$ $(j \in \{1..k\})$ are used to map each object into bit positions to be set to 1 in the filter. In order to test the membership of an object $o \in O$ to the set $\{o_1, \ldots, o_n\}$, the k hash functions can be applied to o: in case at least one maps o to the position of a bit still 0 in the filter, then the membership can be excluded. If o is mapped to bits all set to 1, then o can be one of the objects in the set but we can also be in the presence of a *false positive*. A low False Positive Probability (FPP) is, thus, a quality figure of the filter that has to be minimized via a suitable choice and tuning of the parameters (m, n, k).

In the *standard* Bloom filter usually considered in the recent literature and in the application practice, there are no constraints imposed to the values generated by the k hash functions, so that the same values can be repeatedly generated and less than k bits can be set by the insertion of an object in the filter. In this work, we also consider the variant initially proposed by Bloom [1] in which, for each object, the k hash functions always generate k distinct values and, thus, exactly k bits are set in the filter, as required for the classic superimposed coding [9]. Hence, we will call such variant the *classic* Bloom filter (if the hash functions have disjoint ranges of m/k consecutive bits, this variant corresponds to what has been called *partitioned* Bloom filter in [7]). Notice that the adoption of a classic Bloom filter does not give rise to significant additional computational costs, with respect to a standard Bloom filter, by exploiting the techniques introduced in [7] to avoid hash collisions.

1.1. Background on approximate analysis

After all objects have been inserted, the probability that one bit of the standard Bloom filter is still 0 can be evaluated as $(1 - 1/m)^{kn}$, being the selection of bits to be set with replacement, either with respect to the objects and with respect to the hash functions. If X is a r.v. representing the total number of bits set to 1 in the filter, its expected value is accordingly:

E-mail address: fabio.grandi@unibo.it.

https://doi.org/10.1016/j.ipl.2017.09.004
0020-0190/

$$\mathrm{E}[X] = m\left[1 - \left(1 - \frac{1}{m}\right)^{kn}\right] \tag{1}$$

The main merit figure of the Bloom filter is the False Positive Probability (FPP) that can be computed as the probability that an object non-belonging to $\{o_1, \ldots, o_n\}$ is hashed to only positions with bits set to 1 in the filter. As the bit positions can be chosen with replacement, the probability of a false positive conditioned to a number $X = x$ of bits set to 1 in the standard Bloom filter is given by:

$$\Pr(\mathrm{FP}|X = x) = \left(\frac{x}{m}\right)^k \tag{2}$$

Then the exact value of the FPP can be computed indeed according to the Total Probability theorem as:

$$\begin{aligned}\mathrm{FPP} &= \sum_{x=0}^{m} \Pr(\mathrm{FP}|X = x)\Pr(X = x) \\ &= \sum_{x=0}^{m} f(x)\Pr(\mathrm{FP}|X = x)\end{aligned} \tag{3}$$

where $f(x)$ is the probability mass function of X. Since X can be shown (e.g., via application of the Azuma–Hoeffding inequality [8, Sec. 12.5.3]) to be strongly concentrated around its expected value, a commonly employed approximation is to consider x deterministically equal to $\mathrm{E}[X]$, yielding:

$$\mathrm{FPP} \approx \mathrm{FPP}_{A1} = \left[1 - \left(1 - \frac{1}{m}\right)^{kn}\right]^k \tag{4}$$

Such approximation has been shown in [2] to be highly accurate for large m values with small values of k. Moreover, since $(1 - 1/m)^m \to 1/e$ when m grows, a further asymptotic approximation:

$$\mathrm{FPP} \approx \mathrm{FPP}_{A2} = \left(1 - e^{-kn/m}\right)^k \tag{5}$$

is also commonly used when m is large. FPP_{A2} is minimized when $k = (m/n)\ln 2$, corresponding to one half of the bits set to 1 in the Bloom filter.

No complete analysis of the classic (or partitioned) Bloom filter has been done yet. Kirsch and Mitzenmacher in [7] limit themselves to observe that it tends to have more 1's than the standard Bloom filter and, thus, yields an higher FPP although their asymptotic behavior is the same.

In this paper, we will apply the γ-transform approach described in [5] and that we first introduced in [4] to the analysis both of the standard and of the classic Bloom filters. In this way, in Sec. 2, we will easily derive the exact probability mass function, expected value and variance of the number of bits set to 1, and the FPP of the standard and classic Bloom filters. We will also introduce two iterative schemes for the direct computation of those FPPs and a new accurate non-iterative approximation for the estimation of the FPP of the standard Bloom filter. For small Bloom filters, for which asymptotic approximations are not justified, a comparison between the FPPs of the standard and classic Bloom filters, the new approximation and the old ones can be found in Sec. 3. A Conclusion section will finally close the paper.

2. A new analysis of Bloom filters

In this Section, we exploit the γ-transform approach [4, 5] for the probabilistic characterization of the standard and classic Bloom filters. In a counting experiment where possible outcomes can be selected from a set with cardinality m, the γ-transform $\gamma(y)$ of the probability mass function of the number of outcomes can be evaluated as the probability of selecting outcomes from a subset with cardinality $y \le m$ only. In our case, we can consider as an outcome a bit set to 1 in the filter so that X represents the number of outcomes. Ready-made formulas will then allow us to derive from $\gamma(y)$ the probability mass function of X, the expected value and variance of X.

2.1. Standard Bloom filter

Owing to the physical meaning of the γ-transform recalled above [5, Th. 3], since in the standard Bloom filter selection of bits to be set to 1 is with replacement, we have $\gamma_S(y) = (y/m)^{kn}$. Hence, using formulae (6), (13) and (14) of [5], we can derive in a straightforward way from $\gamma_S(y)$ the probability mass function, expected value and variance of X, respectively, as:

$$f_S(x) = \binom{m}{x}\sum_{j=0}^{x}(-1)^j\binom{x}{j}\left(\frac{x-j}{m}\right)^{kn} \tag{6}$$

$$\mathrm{E}[X] = m\left[1 - \left(1 - \frac{1}{m}\right)^{kn}\right] \tag{7}$$

$$\begin{aligned}\sigma_X^2 = {} & m\left(1 - \frac{1}{m}\right)^{kn} \\ & \times\left[1 - m\left(1 - \frac{1}{m}\right)^{kn} + (m-1)\left(1 - \frac{1}{m-1}\right)^{kn}\right]\end{aligned} \tag{8}$$

The probability mass function $f_S(x)$ is the one we first derived in [4] for a particular case of the "set union problem" and agrees with the expressions derived for the standard Bloom filter in [2,3], while $\mathrm{E}[X]$ in (7) is the same as in (1). As far as we know, no derivation of σ_X^2 has been done by other authors. Notice that the explicit knowledge of σ_X^2 is a good indicator for evaluating how the distribution of X is actually concentrated around $\mathrm{E}[X]$ and, thus, of the goodness of the proposed approximations FPP_{A1} and FPP_{A2} (also for small m).

Using (3) with (6) and (2), the exact expression of the FPP for the standard Bloom filter can then be computed as:

$$\mathrm{FPP}_S = \sum_{x=0}^{m}\left(\frac{x}{m}\right)^k\binom{m}{x}\sum_{j=0}^{x}(-1)^j\binom{x}{j}\left(\frac{x-j}{m}\right)^{kn} \tag{9}$$

which agrees with the expressions derived in [2,3] and is a rather complex formula to evaluate.