ブルームフィルタの設計の理論

Hirotsugu Seike

2025 年 12 月 1 日

1　導入

ブルームフィルタとは，Burton H. Bloom が 1970 年の論文「許容誤差を伴うハッシュ符号化における空間と時間のトレードオフ」において提案したデータ構造である [1]. これは, あるデータ x が集合 $\mathcal{D}$ に含まれているかどうかを, 高い確率で判定するためのスキームである. データ x をハッシュ関数 h_1, h_2 (ハッシュ関数の個数 $k = 2$ とした) に入力し, その結果を m で mod し, ビット配列の該当するインデックスのビットを 1 とする (全てのインデックスの初期値は 0). データ y についても同様な作業を行う. m が十分に大きい場合, ハッシュ関数の出力が一様乱数と見做せることから, その衝突確率は十分に小さくできる. 集合 $\mathcal{D} = \{x, y\}$ に含まれる全ての要素について, 上記プロセスを実行し, 得られた m ビットのビット配列がブルームフィルタである (図 1 を参照).

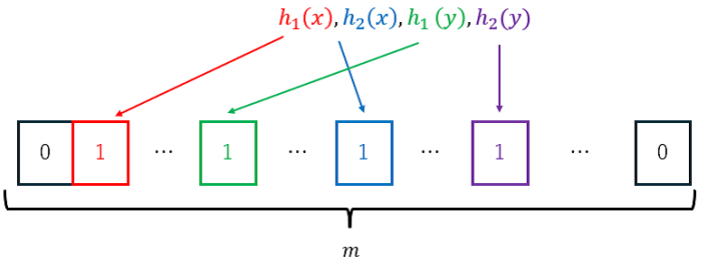

図 1: m ビット配列であるブルームフィルタ例. 採用するハッシュ関数の個数 $k = 2$. $|\mathcal{D}| = 2$.

集合 $\mathcal{D}$ で定義したブルームフィルタを用いて, 集合 $\mathcal{D}$ に含まれない入力 z について, 同様に k 個のハッシュを計算し, ブルームフィルタで 1 になっていないインデックスのビットが一つでも 1 となれば, $z \notin \mathcal{D}$ と判定できる (この計算は, 木構造を使わず, 配列で実装するため, 計算量は $O(k)$ であり高速であることに注意する). また, 運が悪く全ての $h_1(z), h_2(z)$ で計算したビットがブルームフィルタに含まれる事象の確率を偽陽性確率と呼ぶ. ブルームフィルタの構成方法から, 偽陰性 (要素 $z' \in \mathcal{D}$ であるにも関わらず, $z' \notin \mathcal{D}$ となるケース) が存在しないことに注意する.

2　偽陽性確率を最小にするブルームフィルタの設計

サイズ n の集合 $\mathcal{D}$ から作成された m ビット配列のブルームフィルタに対して, $z \notin \mathcal{D}$ である z の偽陽性確率 P_F を最小にするためのハッシュ関数の個数 k を決定する方法について考える. 理想的なハッシュ関数の場合, 出力は一様分布に従うと見做せるため, P_F は次式で与えられる.

$$P_F = \left(1 - \left(1 - \frac{1}{m}\right)^{kn}\right)^k. \tag{1}$$

$(1-1/m)^{kn}$ は, z のハッシュの一つである $h(z) \mod m$ に該当するインデックスのビットが, ブルームフィルタで 0 となっている確率である. つまり, その余事象である $1-(1-1/m)^{kn}$ は, 該当ビットが 1 である確率である. つまり, それを k 乗した値である (1) 式は, 全てのハッシュがブルームフィルタに含まれる確率を表す. m が十分に大きい場合, $(1-1/m)^{-m} \approx (1+1/m)^m \approx e$ であるため, P_F は次式で近似できる.

$$P_F = \left(1 - e^{-kn/m}\right)^k \equiv \left(1 - e^{-ak}\right)^k. \tag{2}$$

ここで, $a = n/m$ である. 上式を k で微分した値は次式で表せる.

$$\frac{\partial P_F}{\partial k} = \left(1 - e^{-ak}\right)^k \cdot \left(\log(1 - e^{-ak}) + \frac{ak \cdot e^{-ak}}{1 - e^{-ak}}\right). \tag{3}$$

(3) 式は, $k = \log(2)/a = m/n \cdot \log(2)$ の時 0 となり, P_F が最小値 $P_{F,min}$ を取る.

$$P_{F,\min} = \left(\frac{1}{2}\right)^k \approx 0.6185^{m/n}. \tag{4}$$

次に, 上記 n, m, k の関係をプロットしたグラフを載せる (k は $m/n \cdot \log(2)$ を四捨五入した値を代入している).

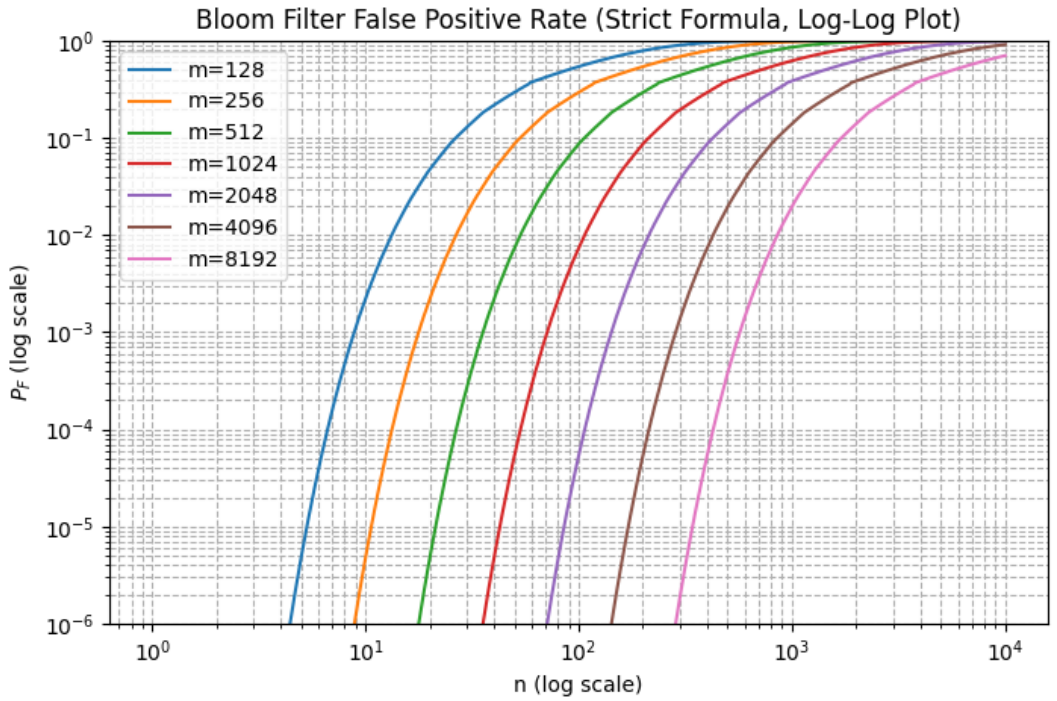

図 2: 偽陽性確率のプロット.

参考文献

[1] Bloom, B. H., "Space/Time Trade-Offs in Hash Coding with Allowable Errors," *Communications of the ACM*, vol. 13, no. 7, pp. 422–426, 1970.