A small approximately min-wise independent family of hash functions

Piotr Indyk*

Stanford University
indyk@cs.stanford.edu

November 21, 1998

Abstract

In this paper we give a construction of a small approximately min-wise independent family of hash functions. The number of bits needed to represent each function is $O(\log n \cdot \log 1/\epsilon)$. This construction gives a solution to the main open problem of [2].

1 Introduction

A family of functions $\mathcal{H} \subset [n] \to [n]$ (where $[n] = \{0 \dots n-1\}$) is called *ϵ-min-wise independent* if for any $X \subset [n]$ and $x \in [n] - X$ we have

$$\Pr_{h \in \mathcal{H}}[h(x) < \min h(X)] = \frac{1}{|X|+1}(1 \pm \epsilon)^1$$

This definition can be generalized to the case when $|X|$ is restricted to be smaller than a prespecified bound s. Such families (restricted to the case when all functions from $\mathcal{H}$ are permutations) were introduced and investigated in [2] and independently earlier in [6] (cf. [7]). The motivation for studying such families is to reduce amount of randomness used by algorithms [6, 2, 3]. In particular (as pointed out in [2]) they have immediate application to efficient detection of similar documents in large text corpora. More specifically, they are crucial components in the following algorithm (introduced in [4] and independently in [5]). Assume that each document is represented as a subset of some universe $U = [n]$ (eg. the set of all words). It was observed in [4] and [5] that the similarity between two documents A and B defined as

$$r(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

captures well the informal notion of being "roughly the same". The value of $r(A, B)$ can be estimated by precomputing for each set A its *sketch* $\overline{A} = \min h_1(A), \dots, \min h_k(A)$, where the functions h_i are chosen independently and uniformly at random from some family $\mathcal{H} \subset [n] \to [n]$. It can be shown [1, 2] that if $\mathcal{H}$ is ϵ-min-wise independent, then for each $i = 1 \dots k$ we have

$$\Pr[h_i(A) = h_i(B)] = r(A, B) \pm \epsilon.$$

Thus by comparing $\overline{A}$ and $\overline{B}$ we get a good estimation of $r(A, B)$. The benefit of using the sketch is that its size is much smaller than the size of the document, which enables to decrease significantly the time and space requirements of the procedure. The above algorithm has been implemented in the Altavista search engine.

In order for this scheme to be useful, it should be possible to represent each function h from $\mathcal{H}$ using small number of bits, as well as evaluate h quickly. In particular, these requirements exclude the family of all functions, as it they require $n \log n$ bits to represent each function, which is prohibitively large (in practice $n = 2^{64}$). In the implementation integrated with Altavista the set $\mathcal{H}$ was chosen to be the a pairwise independent family of hash functions. However, as shown in [2], for such families the error ϵ can be as large as $\log n$ (although it is constant for random sets [2]).

In this paper we address this problem and prove the following theorem.

THEOREM 1.1. *There exist constants $c, c' > 1$ such that for any $\epsilon > 0$ and $s \le \epsilon n/c$ any $c' \log 1/\epsilon$-wise independent family of functions is ϵ-min-wise independent.*

Let $l = c' \log 1/\epsilon$. By applying the above theorem to the family of $l-1$-degree polynomials over $GF(n)$, (which is known to be l-wise independent if n is a prime), we obtain an approximately min-wise independent family of size $n^{O(\log 1/\epsilon)}$. Moreover, each function from that family can be evaluated using $O(\log 1/\epsilon)$

*This work was supported by Stanford Graduate Fellowship and NSF Award CCR-9357849, with matching funds from IBM, Mitsubishi, Schlumberger Foundation, Shell Foundation, and Xerox Corporation.

[1] Here and in the rest of this paper $\pm\epsilon$ denotes a constant δ such that $|\delta| \le \epsilon$.

arithmetic operations by a very simple algorithm. These parameters and simplicity of the construction make it attractive option for the above applications.

Very recently we learned that a similar result has been obtained by Mulmuley [6] (cf. [7], page 399). However, his proof technique seem to require the family of hash functions to be $\Omega(\frac{1}{\epsilon})$-independent, thus for small ϵ the bound implied by our theorem is much smaller.

2 The proof

The basic idea of the proof is to show that the uniform distribution over $\mathcal{H}$ resembles closely the uniform distribution over all functions $[n] \rightarrow [n]$. This implies the desired result, as it is easy to see that for $s \leq \epsilon n/c$ the set of all functions is ϵ-min-wise independent. To show this relation, we consider the event that the image of a fixed subset of a domain is disjoint from a fixed subset of the range. In the language of urn models, this corresponds to the event in which none of the balls thrown goes to a given set of urns. We show that that if the balls are thrown using $O(\ln 1/\epsilon)$-wise independent distribution, then the probability of this event differs from the corresponding probability under uniform distribution by at most ϵ if the number of urns is "small", or even less when the number of urns is "large". In the first case, we use the inclusion-exclusion formula; in the second we exploit the l-th moment method. Finally, we notice that x is a minimum under a function h iff none of the other elements of X fall into the the interval $[0 \ldots h(x)]$. By summing up the probabilities over all values of $h(x)$ we obtain the desired estimation.

The formal statement of the above observations is as follows. Let $k = |X|+1$. For any $i \in [n]$ let A_i denote the event "$h(x) \cap [0 \ldots i] = \emptyset$". Define $E_i = \frac{i+1}{n}(k-1)$; notice that E_i is the expected number of elements from X falling into $[0 \ldots i]$. Let $\Pr_l[\cdot]$ denote any l-wise independent probability measure over $[n] \rightarrow [n]$. Then

Lemma 2.1. *Let i be such that $E_i \leq \frac{l-1}{2e}$ and let $l \geq \log 2/\epsilon$. Then $\Pr_l(A_i) = \Pr(A_i) \pm 2\epsilon$.*

Lemma 2.2. *Let l' be even. Then*

$$\Pr_{l'}(A_i) \leq 48 \left(\frac{6l'}{E_i}\right)^{(l'-1)/2}.$$

Assuming the two lemma hold, we give the proof of the Theorem 1.1. More specifically, we show the following Lemma which can be easily shown to imply Theorem 1.1.

Lemma 2.3. *Let $l' = 2\log 1/\epsilon + 3$, $t = 12l'$ and $l = et+1$ (assume l is even). Then*

$$\Pr_{l+1}[\min h(X) > h(x)] = \Pr[\min h(X) > h(x)] \pm \frac{98\epsilon t}{k-1}$$

Proof. (of Lemma 2.3)

$$\begin{aligned}
&\Pr_{l+1}[\min h(X) > h(x)] \\
&= \sum_{i=0}^{n-1} \Pr_{l+1}[h(x) = i, h(X) \cap [0 \ldots i] = \emptyset] \\
&= \sum_{i=0}^{n-1} \frac{1}{n} \Pr_{l+1}[A_i | h(x) = i] \\
&= \frac{1}{n} \sum_{i=0}^{\frac{nt}{k-1}-1} \Pr_{l+1}[A_i | h(x) = i] + \frac{1}{n} \sum_{i=\frac{nt}{k-1}}^{n-1} \Pr_{l}[A_i | h(x) = i] \\
&= P_1 + P_2
\end{aligned}$$

As the distribution $\Pr_{l+1}[\cdot | h(x) = i]$ is l-wise independent, the component P_1 can be estimated easily from Lemma 2.1 as

$$\begin{aligned}
P_1 &= \frac{1}{n} \sum_{i=0}^{\frac{nt}{k-1}-1} (\Pr[A_i] \pm 2\epsilon) \\
&= \frac{1}{n} \sum_{i=0}^{\frac{nt}{k-1}-1} \Pr[A_i] \pm \frac{2\epsilon t}{k-1}
\end{aligned}$$

Now we estimate the component P_2. From Lemma 2.2 applied to the probability space $\Pr_{l'+1}[\cdot | h(x) = i]$ we have

$$\begin{aligned}
P_2 &\leq \frac{1}{n} \sum_{i=\frac{nt}{k-1}}^{n-1} 48 \left(\frac{6l'}{\frac{(i+1)k-1}{n}}\right)^{\frac{l'-1}{2}} \\
&\leq \frac{1}{n} 48 \left(\frac{6l'n}{k-1}\right)^{\frac{l'-1}{2}} \sum_{i=\frac{nt}{k-1}}^{\infty} \frac{1}{(i+1)^{(l'-1)/2}} \\
&\leq \frac{1}{n} 48 \left(\frac{6l'n}{k-1}\right)^{\frac{l'-1}{2}} \frac{1}{\left(\frac{n}{k-1}t\right)^{(l'-3)/2}} \\
&\leq \frac{t}{k-1} 48 \left(\frac{6l'}{t}\right)^{(l'-3)/2}
\end{aligned}$$

As we assumed $t = 12l'$, we have $P_2 \leq 48\frac{t}{k-1}\frac{1}{2^{(l'-3)/2}} = 48\frac{t}{k-1}48\epsilon$. Thus

$$\Pr_{l+1}[\min h(X) > h(x)] = P_1 + P_2 = \frac{1}{n} \sum_{i=0}^{\frac{nt}{k-1}-1} \Pr[A_i] \pm \frac{50\epsilon t}{k-1}$$

On the other hand (from the analysis as above for P_2) we know that

$$\Pr[\min h(X) > h(x)] = \frac{1}{n} \sum_{i=0}^{\frac{nt}{k-1}-1} \Pr[A_i] \pm 48\frac{t}{k-1}\epsilon$$

The lemma follows.

We now prove the two auxiliary Lemmas.

Proof. (of Lemma 2.1) We use the inclusion-exclusion formula. Specifically

$$\begin{aligned} & \Pr_l[h(X) \cap [0\dots i] = \emptyset] \\ = & \ 1 + \sum_{j=1}^{l-2} (-1)^j \sum_{X' \subset X, |X'|=j} \Pr_l[h(X') \subset [0\dots i]] + \delta \end{aligned}$$

where

$$|\delta| = \sum_{j=l-1}^{l} \sum_{X' \subset X, |X'|=j} \Pr_l[h(X') \subset [0\dots i]]$$

From the l-way independence we known that

$$\sum_{X' \subset X, |X'|=j} \Pr[h(X') \subset [0\dots i]] = \binom{k-1}{j} \left(\frac{i+1}{n}\right)^j$$

Thus

$$\Pr_l[h(X) \cap [0\dots i] = \emptyset] = \sum_{j=1}^{l-2} (-1)^j \binom{k-1}{j} \left(\frac{i+1}{n}\right)^j + \delta$$

such that '

$$\begin{aligned} |\delta| & \leq 2\binom{k-1}{l-1}\left(\frac{i+1}{n}\right)^{l-1} \\ & \leq 2\left(\frac{e(k-1)}{l-1}\right)^{i-1}\left(\frac{i+1}{n}\right)^{l-1} \\ & \leq 2\left(\frac{e(k-1)^{i+1}}{(l-1)n}\right)^{l-1} \\ & \leq 2\left(\frac{eE_i}{l-1}\right)^{l-1} \\ & \leq 2\left(\frac{1}{2}\right)^{\log 2/\epsilon} = \epsilon \end{aligned}$$

On the other hand by the same argument as above

$$\Pr[h(X) \cap [0\dots i] = \emptyset] = \sum_{j=1}^{l-2} (-1)^j \binom{k-1}{j} \left(\frac{i+1}{n}\right)^j + \delta'$$

where $|\delta| \leq \epsilon$. The Lemma follows.

Proof. (of Lemma 2.2) We use the l'-th moment method (we need to assume that l' is even). More specifically, let Z_j, $j = 1\dots k$ be indicator variable which is equal to 1 iff the ith element of X falls into $[0\dots i]$ and is 0 otherwise. Moreover define $Z = \sum_j Z_j$. We will use the inequality

$$Pr_{l'}(A_i) \leq \Pr_l[|Z - E_i| \geq E_i] \leq \frac{\Delta_i^{l'}}{E_i^{l'}}$$

where $\Delta_i^{l'}$ is the l'-th central moment of Z. As its value is equal to the l'-th moment of Z under uniform distribution, we will estimate it by using Chernoff bound, ie. the inequality

$$\Pr[|Z - E_i| \geq \epsilon E_i] \leq 2e^{-\epsilon^2/3E_i}$$

We can write

$$\begin{aligned} E[(Z-E_i)^{l'}] & \leq 2\sum_{j=1}^{\infty} j^{l'} \cdot 2e^{-\frac{j^2}{3E_i^2}E_i} \\ & \leq 4(3E_i)^{(l'+1)/2} \sum_{s=1}^{\infty} s^{l'} e^{-s^2} \end{aligned}$$

It is easy to show that

$$\sum_{s=1}^{\infty} s^{l'} e^{-s^2} \leq 2(2l')^{(l'+1)/2}.$$

Then $\Delta_i^{l'} \leq 8(6l'E_i)^{(l'+1)/2}$. Thus

$$Pr_{l'}(A_i) \leq \frac{8(6l'E_i)^{(l'+1)/2}}{E_i^{(l'-1)/2}} = 48l'\left(\frac{6l'}{E_i}\right)^{(l'-1)/2}$$

Acknowledgements: The author would like to thank Moses Charikar for telling him that Raimund Seidel told him about the theorem by Mulmuley.

References

[1] A. Broder, "On the resemblance and containment of documents", SEQUENCES'98, p. 21-29.

[2] A. Broder, M. Charikar, A. Frieze, M. Mitzenmacher, "Min-wise independent permutations", STOC'98.

[3] A. Broder, M. Charikar, M. Mitzenmacher, "A derandomization using pairwise independent permutations", RANDOM'98.

[4] A. Broder, S. Glassman, M. Manasse, G. Zweig, "Syntactic clustering of the Web", WWW6, p. 391-404, 1998.

[5] E. Cohen, "Size-estimation framework with applications to transitive closure and reachability", FOCS'94, pp. 190-200.

[6] K. Mulmuley, "Randomized Geometric Algorithms and Pseudorandom Generators", *Algorithmica* 16(4/5), pp. 450-463 (1996)

[7] K. Mulmuley, "Computational geometry: an introduction through randomized algorithms", Prentice Hall, 1994.