Journal of Algorithms **38,** 84–90 (2001)
doi:10.1006/jagm.2000.1131, available online at http://www.idealibrary.com on IDEAL®

A Small Approximately Min-Wise Independent Family of Hash Functions

Piotr Indyk[1]

Department of Computer Science, Stanford University, Stanford, California 94305
E-mail: indyk@cs.stanford.edu

Received June 7, 1999

In this paper we give a construction of a small approximately min-wise independent family of hash functions, i.e., a family of hash functions such that for any set of arguments X and $x \in X$, the probability that the value of a random function from that family on x will be the smallest among all values of that function on X is roughly $1/|X|$. 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 A. Broder *et al.* (*in* "STOC'98"). © 2001 Academic Press

1. INTRODUCTION

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

$$\Pr_{h\in\mathscr{H}}[h(x) < \min h(X)] = \frac{1}{|X|+1}(1 \pm \epsilon),^{2}$$

where $\Pr_{h\in\mathscr{H}}$ denotes the probability space obtained by choosing h uniformly at random from $\mathscr{H}$. This definition can be generalized to the case when $|X|$ is restricted to be smaller than a prespecified bound s; we will say that such families are (ϵ, s)-min-wise independent. Clearly, the family of all functions from $[n]$ to $[n]$ is approximately min-wise independent, provided that s is small enough so that $h(x)$ is usually unique. In this paper

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

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

0196-6774/01 $35.00

we are interested in construction of approximately min-wise independent families of small size. Such families (restricted to the case when all functions from $\mathscr{H}$ are permutations) were introduced and investigated in [2] and earlier in [7] (cf. [8]). The motivation for studying such families is to reduce amount of randomness used by algorithms [7, 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]; see also [5]). Assume that each document is represented as a subset of some universe $U = [n]$ (e.g., the set of all words). It was observed in [4] 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* $\overline{A} = \min h_1(A), \ldots, \min h_k(A)$, where the functions h_i are chosen independently and uniformly at random from some family $\mathscr{H}$. It can be shown [1, 2] that if $\mathscr{H}$ is ϵ-min-wise independent, then for each $i = 1, \ldots, k$ we have

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

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 one 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 is helpful if we can represent each function h from $\mathscr{H}$ using a small number of bits, as well as evaluate h quickly. In particular, these requirements exclude the family of all functions, as it requires $n \log n$ bits to represent each function, which is prohibitively large. In the implementation integrated with Altavista the set $\mathscr{H}$ was chosen to be 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 that any *l-wise independent* family of hash function is also approximately min-wise independent, for l large enough. Recall that a family $\mathscr{H} = \{h_i : [n] \to [n]\}$ is called *l-wise independent* if for any $X = \{x_1 \cdots x_l\} \subset [n]$, $Y = \{y_1 \cdots y_l\} \subset [n]$, we have $\Pr_{h \in \mathscr{H}}[h(x_i) = y_i, i = 1, \ldots, l] = 1/n^l$. The formal statement of the theorem is as follows.

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

Note that the upper bound on s is necessary even for completely random functions in order to guarantee that the minimum is unique with a sufficient probability (a family of *permutations* would not suffer from this problem).

Let $l = c' \log 1/\epsilon$. By applying the above theorem to the family of polynomials of degree $l-1$ 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 $O(\log 1/\epsilon)$ arithmetic operations by a very simple algorithm. These parameters and simplicity of the construction make it an attractive option for the above applications.

Very recently we learned that a similar result has been obtained by Mulmuley [7] (cf. [8], p. 399). However, his proof technique seems to require the family of hash functions to be $\Omega(\frac{1}{\epsilon})$-wise 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 the family $\mathscr{H}$ as in Theorem 1.1 resembles closely the uniform distribution over all functions from $[n]$ to $[n]$. 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 if the balls are thrown using $O(\log 1/\epsilon)$-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 lth moment method. Finally, we notice that x is a minimum under a function h if and only if none of the other elements of X falls into the interval $[0, h(x)]$. 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 $k = |X|$. For any $i \in [n]$ let A_i denote the event "$h(X) \cap [0, i] = \emptyset$." Define $E_i = \frac{i+1}{n} k$; notice that E_i is the expected number of elements from X falling into $[0, i]$. Let $\Pr_l[\cdot]$ denote any probability measure induced by random choice of a function from any l-wise independent family of functions.

LEMMA 2.1. *Let i be such that $E_i \leq \frac{l-1}{2e}$ and let $l \geq \log(2/\epsilon)$. Then $\Pr_l[A_i] = \Pr[A_i] \pm 2\epsilon$, where $\Pr[\cdot]$ denotes the probability space over all functions h from $[n]$ to $[n]$.*

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

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

Assuming that the two lemmas hold, we give the proof of 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 that l is even). Then*

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

Proof (Proof of Lemma 2.3).

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

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

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

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

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

As we assumed $t = 12l'$, we have

$$P_2 \le 48\frac{t}{k}\frac{1}{2^{(l'-3)/2}} = 48\frac{t}{k}48\epsilon.$$

Thus

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

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

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

The lemma follows. ∎

We now prove the two auxiliary lemmas.

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

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

where

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

From the l-way independence we known that

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

Thus

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

such that

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

On the other hand by the same argument as above

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

where $|\delta| \le \epsilon$. The lemma follows. ▮

Proof (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, $j = 1\cdots k$ be an indicator variable which is equal to 1 iff the ith element of X falls into $[0\cdots i]$ and is 0 otherwise. Moreover define $Z = \sum_j Z_j$. We will use the inequality

$$\Pr_{l'}(A_i) \le \Pr_{l}[|Z-E_i| \ge E_i] \le \frac{\Delta_i^{l'}}{E_i^{l'}},$$

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 the Chernoff bound [6], i.e., the inequality

$$\Pr[|Z-E_i| \ge \epsilon E_i] \le 2e^{-\epsilon^2/3E_i}.$$

We can write

$$E[(Z-E_i)^{l'}] \le 2\sum_{j=1}^{\infty} j^{l'} \cdot 2e^{-\frac{j^2}{3E_i^2}E_i}$$

$$\le 4(3E_i)^{(l'+1)/2}\sum_{s=1}^{\infty} s^{l'} e^{-s^2}.$$

It is easy to show that

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

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

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

ACKNOWLEDGMENT

The author thanks 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, *in* “SEQUENCES’98,” pp. 21–29.
2. A. Broder, M. Charikar, A. Frieze, and M. Mitzenmacher, Min-wise independent permutations, *in* “STOC’98.”
3. A. Broder, M. Charikar, and M. Mitzenmacher, A derandomization using min-wise independent permutations, *in* “RANDOM’98.”
4. A. Broder, S. Glassman, M. Manasse, and G. Zweig, Syntactic clustering of the Web, *in* “WWW6,” pp. 391–404, 1998.
5. E. Cohen, Size-estimation framework with applications to transitive closure and reachability, *in* “FOCS’94,” pp. 190–200.
6. R. Motwani and P. Raghavan, “Randomized Algorithms,” Cambridge Univ. Press, Cambridge, UK, 1995.
7. K. Mulmuley, Randomized geometric algorithms and pseudorandom generators, *Algorithmica* **16**, Nos. 4/5 (1996), 450–463.
8. K. Mulmuley, “Computational Geometry: An Introduction through Randomized Algorithms,” Prentice Hall, New York, 1994.