Lecture Notes: Flajolet-Martin Sketch

Yufei Tao
Chinese University of Hong Kong
taoyf@cse.cuhk.edu.hk

12 Feb, 2012

1 Distinct element counting problem

Let S be a *multi-set* of N integers, namely, two elements of S may be identical. Each integer is in the range of $[0, D]$ where D is some polynomial of N. The *distinct element counting problem* is to find out exactly how many distinct elements there are in S. We will use F to denote the answer. For example, given $S = \{1, 5, 10, 5, 15, 1\}$, $F = 4$.

Clearly, using $O(N)$ words of space, the problem can be solved easily in $O(N \log N)$ time by sorting, or $O(N)$ expected time with hashing. In many applications, however, the amount of space at our disposal can be much smaller. In this lecture, we consider that we are allowed only $O(\log N)$ *bits*. Hence, our goal is to obtain an approximate answer $\tilde{F}$ whose accuracy has a probabilistic guarantee.

We will learn a structure proposed by Flajolet and Martin [2] that can achieve this purpose by seeing each element of S only *once*. We will name the structure the *FM-sketch* after the inventors. Let w be the smallest integer such that $2^w \geq N$, that is, $\lceil w = \log N \rceil$. For simplicity, we assume that there is an ideal hash function h which maps each integer $k \in S$ independently to a hash value $h(k)$ that is distributed uniformly in $[0, 2^w - 1]$.

2 FM-sketch

Each integer k in $[0, 2^w - 1]$ can be represented with w bits. We will use z_k to denote the number of leading 0's (counting from the left) in the binary form of the hash value $h(k)$ of k. For example, if $w = 5$ and $h(k) = 6 = (00110)_2$, then $z_k = 2$ because there are two 0's before the leftmost 1. The FM sketch is simply an integer Z defined as:

$$Z = \max_{k \in S} z_k. \tag{1}$$

Clearly, Z can be obtained by seeing each element k once: simply calculate z_k, update Z accordingly, and then discard k. Note that the z_k of all $k \in S$ are independent. Also obvious is the fact that Z can be stored in $w = O(\log N)$ bits. After Z has been computed, we simply return

$$\tilde{F} = 2^Z$$

as our approximate answer.

3 Analysis

This section will prove the following property of the FM sketch:

Proposition 1. *For any integer $c > 3$, the probability that $\frac{1}{c} \leq \frac{\tilde{F}}{F} \leq c$ is at least $1 - \frac{3}{c}$.*

Our proof is based on [1]. We say that our algorithm is *correct* if $\frac{1}{c} \leq \frac{\tilde{F}}{F} \leq c$ (i.e., our estimate $\tilde{F}$ is off by at most a factor of c, from either above or below). The above proposition indicates that our algorithm is correct with at least a constant probability $1 - \frac{3}{c} > 0$.

Lemma 1. *For any integer* $r \in [0, w]$, $\mathbf{Pr}[z_k \geq r] = \frac{1}{2^r}$.

Proof. Note that $z_k \geq r$ means that the hash value $h(k)$ of k is between $\underbrace{0...0}_{r}\underbrace{0...0}_{w-r}$ and $\underbrace{0...0}_{r}\underbrace{1...1}_{w-r}$, namely, between 0 and $2^{w-r} - 1$. Remember that $h(k)$ is uniformly distributed from 0 to $2^w - 1$. Hence:

$$\mathbf{Pr}[z_k \geq r] = \frac{2^{w-r}}{2^w} = \frac{1}{2^r}.$$

□

Let us fix an r. For each $k \in S$, define:

$$x_k(r) = \begin{cases} 1 & \text{if } z_k \geq r \\ 0 & \text{otherwise} \end{cases}$$

By Lemma 1, we know that $x_k(r)$ takes 1 with probability $1/2^r$. Hence:

$$\mathbf{E}[x_k(r)] = 1/2^r \tag{2}$$

$$\mathbf{var}[x_k(r)] = \frac{1}{2^r}\left(1 - \frac{1}{2^r}\right) \tag{3}$$

Also define:

$$X(r) = \sum_{\text{distinct } k \in S} x_k(r).$$

Let:

$$\begin{aligned} r_1 &= \text{the smallest } r \text{ such that } 2^r > cF \\ r_2 &= \text{the smallest } r \text{ such that } 2^r \geq \frac{F}{c} \end{aligned}$$

Lemma 2. *Our algorithm is correct if* $X(r_1) = 0$ *and* $X(r_2) \neq 0$.

Proof. Our algorithm is correct if Z as given in (1) satisfies $r_2 \leq Z < r_1$, due to the definitions of r_1 and r_2. If $X(r_1) = 0$, it means that no $k \in S$ gives an $z_k \geq r_1$; this implies $Z < r_1$ (see again (1)). Likewise, if $X(r_2) \neq 0$, it means that at least one $k \in S$ gives an $z_k \geq r_2$; this implies $Z \geq r_2$. □

Next, we will prove that the probability of having "$X(r_1) = 0$ and $X(r_2) \neq 0$" is at least $1 - 3/c$. Towards this, we will consider the complements of these two events, namely: $X(r_1) \geq 1$ and $X(r_2) = 0$. We will prove that $X(r_1) \geq 1$ can happen with probability at most $1/c$, whereas $X(r_2) = 0$ can happen with probability at most $2/c$. then it follows from the union bound that the probability of *at least* one of the two events happening is at most $3/c$. This is sufficient for establishing Proposition 1.

Lemma 3. $\mathbf{Pr}[X(r_1) \geq 1] < 1/c$.

Proof.

$$\begin{aligned}
\mathbf{E}[X(r_1)] &= \sum_{\text{distinct } k \in S} \mathbf{E}[x_k(r_1)] \\
\text{(by (2))} &= F/2^{r_1} \\
\text{(by definition of } r_1) &< 1/c.
\end{aligned}$$

Hence, by Markov inequality, we have:

$$\mathbf{Pr}[X(r_1) \geq 1] \leq \mathbf{E}[X(r_1)] < 1/c.$$

□

Lemma 4. $\mathbf{Pr}[X(r_2) = 0] < 2/c$.

Proof. Same as the proof of the previous lemma, we obtain:

$$\mathbf{E}[X(r_2)] = F/2^{r_2}$$

As $X(r_2)$ is the sum of F independent variables, each of which has variance $\frac{1}{2^r}(1 - \frac{1}{2^r})$ (see Equation 3), we know:

$$\mathbf{var}[X(r_2)] = \frac{F}{2^{r_2}}\left(1 - \frac{1}{2^{r_2}}\right) < \frac{F}{2^{r_2}}.$$

Thus:

$$\begin{aligned}
\mathbf{Pr}[X(r_2) = 0] &= \mathbf{Pr}\big[X(r_2) - \mathbf{E}[X(r_2)] = \mathbf{E}[X(r_2)]\big] \\
&\leq \mathbf{Pr}\Big[\big|X(r_2) - \mathbf{E}[X(r_2)]\big| = \mathbf{E}[X(r_2)]\Big] \\
&\leq \mathbf{Pr}\Big[\big|X(r_2) - \mathbf{E}[X(r_2)]\big| \geq \mathbf{E}[X(r_2)]\Big] \\
\text{(by Chebyshev inequality)} &\leq \frac{\mathbf{var}[X(r_2)]}{(\mathbf{E}[X(r_2)])^2} \\
&< \frac{F/2^{r_2}}{(F/2^{r_2})^2} \\
&= \frac{2^{r_2}}{F}
\end{aligned}$$

From the definition of r_2, we know that $2^{r_2} < 2F/c$ (otherwise, r_2 would not be the *smallest* r satisfying $2^r \geq F/c$). Combining this with the above gives $\mathbf{Pr}[X(r_2) = 0] < 2/c$. □

4 Boosting the success probability

Proposition 1 shows that our estimate $\tilde{F}$ is accurate up to a factor $c > 3$ with probability at least $1 - 3/c$. The success probability $1 - 3/c$ does not look very impressive: ideally, we would like to be able to succeed with a probability arbitrarily close to 1, namely, $1 - \delta$ where $\delta > 0$ can be arbitrarily small. It turns out that we are able to achieve this with a simple median trick for $c > 6$.

Let us build s independent FM-sketches, each of which is constructed as explained in Section 2. The value of s will be determined later. From each FM-sketch, we obtain an estimate $\tilde{F}_i$ $(1 \leq i \leq s)$ of F. We determine our final estimate $\tilde{F}$ as the median of $\tilde{F}_1, ..., \tilde{F}_s$. Now we prove that this trick really works:

Theorem 1. *For each constant $c > 6$, there is an $s = O(\log \frac{1}{\delta})$ ensuring that $\frac{F}{c} \le \tilde{F} \le cF$ happens with probability at least $1 - \delta$.*

Proof. For each $i \in [1, s]$, define $x_i = 0$ if $\tilde{F}_i \in [F/c, cF]$, or 1 otherwise. From Proposition 1, we know that $\mathbf{Pr}[x_i = 1]$ is at most $\rho = 3/c < 1/2$. Clearly, $\mathbf{E}[x_i] = \rho$. Let

$$X = \sum_{i=1}^{s} x_i.$$

Hence:

$$\mathbf{E}[X] = s\rho.$$

If $X < s/2$, then $\frac{F}{c} \le \tilde{F} \le cF$ definitely holds. To see this, consider $\tilde{F} > cF$. Since $\tilde{F}$ is the median of $\tilde{F}_1, ..., \tilde{F}_s$, it follows that at least $s/2$ of these estimates are above cF, contradicting $X < s/2$. Likewise, $\tilde{F}$ cannot be smaller than F/c either.

We will show that $X < s/2$ happens with probability at least $1 - \delta$. Towards this, we argue that the complement event $X \ge s/2$ happens with probability at most δ. As $x_1, ..., x_s$ are independent, we have:

$$\begin{aligned}
\mathbf{Pr}[X \ge s/2] &= \mathbf{Pr}[X - \mathbf{E}[X] \ge s/2 - \mathbf{E}[X]] \\
(\text{as } \mathbf{E}[X] = s\rho < s/2) &\le \mathbf{Pr}[|X - \mathbf{E}[X]| \ge s/2 - \mathbf{E}[X]] \\
&= \mathbf{Pr}[|X - \mathbf{E}[X]| \ge s/2 - s\rho] \\
&= \mathbf{Pr}\left[|X - \mathbf{E}[X]| \ge \frac{1/2 - \rho}{\rho} \cdot s\rho\right] \\
(\text{by Chernoff bound}) &\le 2e^{-\frac{(1/2-\rho)^2}{3\rho^2} s\rho} \\
&= 2e^{-\frac{s(1/2-\rho)^2}{3\rho}}
\end{aligned}$$

To make the above at most δ, we need

$$s \ge \frac{3\rho}{(1/2 - \rho)^2} \ln \frac{2}{\delta}.$$

Hence. setting $s = \lceil \frac{3\rho}{(1/2-\rho)^2} \ln \frac{2}{\delta} \rceil = O(\log \frac{1}{\delta})$ fulfills the requirement. □

References

[1] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences (JCSS)*, 58(1):137–147, 1999.

[2] P. Flajolet and G. N. Martin. Probabilistic counting. In *Proceedings of Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 76–82, 1983.