Algorithms Seminar 2002–2004,
F. Chyzak (ed.), INRIA, (2005), pp. 93–96.

Available online at the URL
http://algo.inria.fr/seminars/.

A Probabilistic Counting Algorithm

Marianne Durand
Algorithms Project, INRIA (France)

December 9, 2002

Summary by Pierre Nicodème

Abstract

This talk[1] (a joint work with Philippe Flajolet) presents an algorithm to approximate count the number of different words in very large sets or texts (in the range of billions of bytes) and its analysis. When using an auxiliary memory of m bytes, the accuracy is of the order $1/\sqrt{m}$. The analysis of this new algorithm relies on asymptotic Depoissonization techniques.

1. Introduction

The problem addressed by this work is how to estimate the number of distinct elements in a large collection of data with the following requirements while: doing a single pass on the data; using a small amount of memory; doing only a few computations; doing no assumptions about the distribution of the data.

Applications of such a problem are data mining optimizations and routers programmation.

2. Summary of Some Algorithms

2.1. **Previous work.** In the following, the input words are considered as elements of $[0, 1]$ (take the 2-adic value of each word represented as a bit-string). We consider the number of input items N, the standard deviation σ of the rate of error done on the counting and the space S needed by the algorithm. In any case a hashing of the data is done before performing the algorithms. The algorithms precedently used for these aims may be classified as:

- **(adaptive) hashing schemes** [3]
  - ∗ algorithm: hash the values in $[0, 1/2]$ in a table; when the number of collisions exceeds a given value γ, skip to a different and smaller interval (by instance $[1/2, 3/4]$); return a function of the number of collisions.
  - ∗ parameters: $\sigma = 1.5/\sqrt{S}$; assumptions on the size of the data; unstable with respect to the order of arrival of the data;
- **adaptive sampling** [4]
  - ∗ algorithm: maintain one bucket of size m; when the bucket overflows, throw away all data beginning by a 1; filter out incoming data beginning by a 1; repeat the process by filtering with 00 and so on; return a function of the number of iterations;
  - ∗ parameters: $\sigma = 6.7/\sqrt{m}$ (unprecise algorithm);
- **probabilistic counting** [6]

[1] The results presented here and recent improvements will be presented at the ESA 2003 Symposium [2].

- ∗ algorithm: let $\rho(w)$ be the position of the first 1 in w; (by instance, $\rho(00010111) = 4$); set up the bits corresponding to $\rho(w)$ for all w in a bit-map. Let k_{max} be the first bit equal to 0 in this bit-map. The estimator is $2^{ck_{max}}$ for a given c.
- ∗ parameters: $\sigma = 0.78/\sqrt{m}$, $S = m \times \log_2 N$;

See also [1].

2.2. **The new algorithm of Durand and Flajolet.** This algorithm uses a technique of *maximum-based probabilistic counting*. It has the following features:

– algorithm: send the data to $m = 2^b$ different buckets, according to the value of their b first bits. For each bucket i compute the maximum

$$M^{(i)} = \max(\{\rho(\mathrm{suf}(w)); w \text{ is hashed in bucket } i\}),$$

where $\mathrm{suf}(w)$ is the suffixe of w starting at position $b + 1$.

$$\text{Return} \quad E = \alpha_m m \times 2^{\frac{1}{m}\sum M^{(i)}}, \tag{1}$$

$$\text{where} \quad \alpha_m = \left(\Gamma(-1/m)\frac{1 - 2^{1/m}}{\log 2}\right)^{-m}, \qquad \Gamma(s) = \frac{1}{s}\int_0^\infty e^{-t}t^s dt; \tag{2}$$

– parameters: $\sigma = 1.3/\sqrt{m}$; memory $S = m \times \log\log\max(\{M^{(i)}\})$;
– remarks: the algorithm is independent of the repetitions and need very few computations. It is only necessary to maintain one value of size $O(\log\log N)$ for each bucket and not a bitmap of size $O(\log N)$ as in probabilistic counting.

3. **Analysis**

As frequently observed, the analysis is easier when Poissonization-Depoissonization is used. The steps of the analysis therefore are.

1. Compute the generating function $F(z) = \sum f_n z^n$, where f_n is the estimator of number of different items when exactly n are read by the algorithm.
2. "Poissonize" the system by considering that the number of items read by the algorithm is a random number following a Poisson distribution of parameter λ. Asymptotically, when $\lambda = n$, one expects the Poisson model to reflect corresponding properties of the fixed-n model (note that for large λ, the Poisson law is concentrated near its mean).
   During this step, compute
   $$\widetilde{F}(z) = \sum f_n \frac{z^n}{n!} e^{-z} = \sum \widetilde{F}_n z^n.$$
3. Compute the Mellin transform $f^\star(s)$ of $\widetilde{F}(z)$. The expansions of $f^\star(s)$ in the neighborhood of its singularities give the asymptotic value of $\widetilde{F}_n$.
4. Prove by depoissonization that, asymptotically, $f_n \sim \widetilde{F}_n$.

3.1. **Getting the basic generating function.** We are interested here to the statistics of the estimator

$$Z = E/\alpha_m = m \times 2^{\frac{1}{m}\sum_i M^{(i)}}.$$

Considering one bucket that receives ν elements, the random variable M is the maximum of ν random variables Y that are independent and geometrically distributed according to $\mathbf{P}(Y \geq k) = 1/2^k$.

Therefore we have

$$\mathbf{P}_\nu(M \le k) = \left(1 - \frac{1}{2^k}\right)^\nu, \quad \text{and} \quad \mathbf{P}_\nu(M = k) = \left(1 - \frac{1}{2^k}\right)^\nu - \left(1 - \frac{1}{2^{k-1}}\right)^\nu.$$

This sums up to

$$G(z,u) = \sum_{\nu,k} \mathbf{P}(M=k) u^k \frac{z^\nu}{\nu!} = \sum_k u^k \left(e^{z(1-1/2^k)} - e^{z(1-1/2^{k-1})}\right). \tag{3}$$

Considering now the $m = 2^b$ buckets induces multinomials when distributing elements amongst buckets; therefore $n![z^n]G(z/m,u)^m$ is the probability generating function of $\sum_i M^{(i)}$.

The expressions for the first and second moment of Z are obtained from there by substituting respectively u by $2^{1/m}$ and $2^{2/m}$.

This gives the following lemma.

Lemma 1. *When there are n input items, the expected value and variance of the unnormalized estimator Z are*

$$\mathbf{E}(Z) = mn![z^n]G\left(\frac{z}{m}, 2^{1/m}\right)^m, \quad \textit{and} \tag{4}$$

$$\mathbf{Var}(Z) = m^2 n![z^n]G\left(\frac{z}{m}, 2^{2/m}\right)^m - \left(mn![z^n]G\left(\frac{z}{m}, 2^{1/m}\right)^m\right)^2. \tag{5}$$

3.2. **Poissonization.** If $f(z) = \sum_n f_n z^n/n!$ is the exponential generating function of the expectation of a parameter, the quantity $e^{-\lambda} f(\lambda) = \sum_n f_n e^{-\lambda}\lambda^n/n!$ gives the corresponding generating function under the Poisson model. Therefore the quantities

$$\mathcal{E}_n = mG\left(\frac{n}{m}, 2^{1/m}\right)^m (e^{-n/m})^m \quad \text{and} \quad \mathcal{V}_n = m^2 G\left(\frac{n}{m}, 2^{2/m}\right)^m e^{-n} - \mathcal{E}_n^2 \tag{6}$$

are respectively the mean and the variance of Z when the number of input items follows a Poisson law of rate $\lambda = n$.

We consider in the following the variable $\mathcal{E}_n$.

Using Equations 3 and 6, we can write

$$\mathcal{E}_n = mA(n)^m, \quad \text{where} \quad A(x) = \sum_i \frac{2^i}{m}\left(\phi(x/2^i) - \phi(x/2^{i-1})\right), \text{ and } \phi(x) = e^{-x/m}.$$

The Mellin transform $F^\star(s)$ (see [5, 8]) of a harmonic sum $F(x) = \sum \lambda_k f(\mu_k x)$ is

$$F^\star(s) = f^\star(s) \sum \frac{\lambda_k}{\mu_k^s};$$

this implies that

$$A^\star(s) = \phi^\star(s)(2^s - 1)\frac{2^{1/m}}{1 - 2^{1/m}2^s}.$$

The dominant singularity is at $s = -1/m$ and the corresponding residue is

$$a = m^{-1/m}\Gamma(-1/m)\frac{1 - 2^{1/m}}{\log 2}.$$

The Mellin transfer theorem gives the corresponding contribution $ax^{1/m}$ in the asymptotic expansion of $A(x)$ at infinity. The same techniques apply when considering $\mathcal{V}_n$. These results are summarized in the following lemma.

Lemma 2. *The Poisson mean $\mathcal{E}_n$ and variance $\mathcal{V}_n$ satisfy as $n \to \infty$:*

$$\mathcal{E}_n \sim \left[\left(\Gamma(-1/m)\frac{1-2^{1/m}}{\log 2}\right)^m + \eta_n\right] \times n, \tag{7}$$

$$\mathcal{V}_n \sim \left[\left(\Gamma(-2/m)\frac{1-2^{1/m}}{\log 2}\right)^m - \left(\Gamma(-1/m)\frac{1-2^{1/m}}{\log 2}\right)^{2m} + \kappa_n\right] \times n^2, \tag{8}$$

where $|\eta_n|$ and $|\kappa_n|$ (bounded by 10^{-6}) correspond to "negligible" singularities.

3.3. **Depoissonization.** The asymptotic forms of the first two moments of Z in the fixed-n model can be transferred back from the Poisson model by a method called "analytic depoissonization" by Jacquet and Szpankowski (See [7, 8]). The *values* of an exponential generating function at large arguments are closely related to the asymptotic form of its *coefficients* provided the generating function decays fast enough away from the positive real axis in the complex plane.

We have

$$G(z/m, 2^{1/m}) = e^{z/m} \sum_k 2^{k/m} e^{-z2^{-k}/m} \left(1 - e^{-z2^{-k}/m}\right).$$

Let S_θ be the cone

$$S_\theta = \{z : |\arg z| \le \theta\}, \quad \text{with } |\theta| < \pi/2.$$

There exists a θ such that

1. inside the cone S_θ there holds $e^{-z}G(z/m, 2^{1/m})^m = O(|z|)$, and
2. outside the cone S_θ there exists α such that $G(z/m, 2^{1/m})^m = O(e^{\alpha|z|})$.

This implies the following lemma (proof omitted).

Lemma 3. *The first two moments of the estimator Z are asymptotically equivalent under the Poisson and fixed-n model: $\mathbf{E}(Z) \sim \mathcal{E}_n$, $\mathbf{Var}(Z) \sim \mathcal{V}_n$.*

4. Improved Algorithm

An heuristic improvement consists in truncating the large non-meaningful values of the indicators M. When respectively 0%, 10%, 20% and 30% of the higher values are truncated, computations (for 32-bits words) give $\sigma \times \sqrt{S} = 2.8, 2.4, 2.2$, and 2.5.

Bibliography

[1] Alon (Noga), Matias (Yossi), and Szegedy (Mario). – The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences*, vol. 58, 1999, pp. 137–147.

[2] Durand (Marianne) and Flajolet (Philippe). – Loglog Counting of Large Cardinalities. – To be presented at ESA2003.

[3] Estan (Cristian) and Varghese (George). – New directions in traffic measurement and accounting. In *Proceedings of SIGCOMM 2002*. – ACM Press, 2002. (Also: UCSD technical report CS2002-0699, February, 2002; available electronically.).

[4] Flajolet (Philippe). – On adaptive sampling. *Computing*, vol. 34, 1990, pp. 391–400.

[5] Flajolet (Philippe), Gourdon (Xavier), and Dumas (Philippe). – Mellin transforms and asymptotics : Harmonic sums. *Theoretical Computer Science*, vol. 144, n° 1-2, 1995, pp. 3–58.

[6] Flajolet (Philippe) and Martin (G. Nigel). – Probabilistic counting algorithms for data base applications. *Journal of Computer and System Sciences*, vol. 31, n° 2, 1985, pp. 182–209.

[7] Jacquet (Philippe) and Szpankowski (Wojciech). – Analytical depoissonization and its applications. *Theoretical Computer Science*, vol. 201, n° 1-2, 1998.

[8] Szpankowski (Wojciech). – *Average-Case Analysis of Algorithms on Sequences*. – John Wiley, New York, 2001.