Efficient estimation of the cardinality of large data sets

Philippe Chassaing, Lucas Gerin

Efficient estimation of the cardinality of large data sets

Philippe Chassaing and Lucas Gerin

22 september 2006

Abstract

F.Giroire has recently proposed an algorithm which returns the *approximate* number of distincts elements in a large sequence of words, under strong constraints coming from the analysis of large data bases. His estimation is based on statistical properties of uniform random variables in $[0, 1]$. Here we propose an optimal estimation, using information and estimation theory[1].

1 Introduction

1.1 The problem

The aim of this note is to improve a solution proposed by Giroire [Gir05] to the following problem: consider a sequence $Y = (Y_1, \ldots, Y_N)$ of words (one may think to a sequence of file on a disk, a list of requests, a novel from Skakespeare, *etc*...); we don't make any assumption on the structure of Y, and we want to know the number (usually denoted F_0 in the data base community) of *distinct* elements of this sequence. The reader has to keep in mind the fact that in applications, N and F_0 are unknown but assumed to be very large.

The motivation comes from analysis of large data sets, and especially analysis of internet traffic : certain attacks may be detected at router level, because they generate an unusual number of dictinct connections (see [Fla04]), and then the knowledge of F_0 is of great importance. Most of algorithms use a dictionnary to store every word, so that the memory needed is linear in F_0. Here the size of data sets is huge, making it impossible to store every word,

[1] Version of August 29, 2007.

so that the algorithm should satisfy the two following constraints: it should use constant memory and do only one pass over the data. These constraints are very strong, but on the other hand we allow the algorithm to give only an *estimation* of F_0.

The main idea used in [Gir05], introduced by Flajolet and Martin [FM85], is to transform this problem in a probabilistic one, using hash fuctions.

> A *hash function* is a function $h : \mathcal{C} \to [0,1]$, where $\mathcal{C}$ is a finite set of words (say english language, $\{0,1\}^8$, *etc...*) such that *the image of a typical sequence of words behaves as* a sequence of i.i.d random variables, uniform in $[0,1]$.

This definition is of course somewhat informal, but we will assume, from now on, that, noting $X_i = h(Y_i)$, then the random set $\mathbf{X} = \{X_1, \ldots, X_N\}$ is the realization of F_0 i.i.d. r.v. uniform on $[0,1]$. Existence and construction of *good* hash functions is discussed in [Knu73].

Set $\theta = F_0$ and denote as usually $X_{(1)}$ the smallest X_i, $X_{(2)}$ the second smallest, and so on. The key point is that the information on θ contained in $\{Y_1, \ldots, Y_N\}$ is equivalent to that contained in $(X_{(1)}, \ldots, X_{(\theta)})$.

As a consequence, we are now dealing with a classical statistical problem: given a (small) sample of $(X_1, \ldots, X_\theta)$, i.i.d. r.v., uniform on $[0,1]$, we want to estimate the (large) parameter θ. Denote by M the memory available, that is the number of real numbers that can be stored during the algorithm. One should determine:

1. A way of extracting a M-sample of $\mathbf{X}$ (the M smallest, the M with the longest sequence of zeros in their binary representation, *etc...*).
2. A function $\hat{\xi} : [0,1]^M \to \mathbb{R}$ which approximates θ, when applied to this M-sample.

State of the Art. Flajolet and Martin [FM85] have used these ideas to construct an algorithm based on research of patterns of 0's and 1's in the binary representation of the hashed values $X_1, \ldots, X_\theta$. It has been improved by Durand and Flajolet [DF93]. Bar-Yossef *et alii* [BYJK$^+$02], have proposed 3 performant algorithms, their ideas have been generalized by Giroire [Gir05].

In a different way, Alon, Matias, and Szegedy consider estimation by *moment method*, making implementation proposed in [FM85] easier. For a nice survey about these ideas one may read [Fla04].

1.2 The algorithm

The starting idea in [Gir05] is to use this simple property:

$$\mathbb{E}[X_{(1)}] = \frac{1}{\theta + 1}.$$

Consequently, a naive algorithm would hash every data, compare it to the smallest hashed value already seen, and finally return $1/X_{(1)}$. Unfortunately, $\mathbb{E}[1/X_{(1)}] = \infty$. However, $1/X_{(2)}, 1/X_{(3)} \dots$ have finite expectation. This leads Giroire to propose an algorithm which return a function of $X_{(k)}$, for some k. In order to improve the precision of such an algorithm, one may wish to execute it m times with m different hashing functions, but this would cost too much time. Therefore Giroire uses *stochastic averaging*, introduced in [FM85]: the idea is to *simulate* m different experiments, by dividing $[0, 1]$ in m intervals.

Let us now describe Giroire's algorithm.

Algorithm 1 ([Gir05]).
let k, m be integers.
Set $Z_{(p),i} = \frac{i}{m}$ for all i, p.
for $j = 1$ to N
    $Z_j = h(Y_j)$.
    let i the integer such that Z_j lies in $[\frac{i-1}{m}, \frac{i}{m}[$.
    update the vector $(Z_{(1),i}, \dots, Z_{(k),i})$ of the k smallest values already seen lying in $[\frac{i-1}{m}, \frac{i}{m}[$.
next j.
for all p, i, set $X_{(p),i} = Z_{(p),i} - \frac{i-1}{m}$.
return an estimator $\hat{\xi} = \hat{\xi}(X_{(l),i}; i = 1, \dots, m; l = 1, \dots, k)$.

Of course this algorithm is not completely described, since $\hat{\xi}$ has not been defined yet. We get m vectors in $[0, 1/m]^k$. $X_{(k),i}$ is the k-th smallest hashed

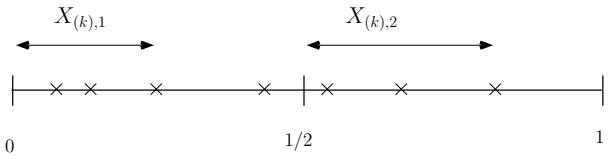

Figure 1: A sample of $(X_{(k),1}, \dots, X_{(k),m})$, with $\theta = 5, m = 2, k = 3$.

value lying in $[\frac{i-1}{m}, \frac{i}{m}]$, shifted by $(i-1)/m$. If less than l values have fallen in the i-th interval, then $X_{(k),i} = 1$. Obviously, Algorithm 1 makes only one

pass over each data Y_i. Memory used by the algorithm is indeed M, if we have chosen $k \cdot m = M$. The estimation returned by the algorithm does not depend on any assumption on the repetitions in the sequence $X_1, \ldots, X_N$.

Giroire [Gir05] proposes 3 estimators ξ_1, ξ_2, ξ_3, using inverse, square root and log respectively. For example,

$$\xi_3 := \left(\frac{\Gamma(k-1/m)}{\Gamma(k)}\right)^{-m} . e^{-\frac{1}{m}\sum_{i=1}^{m} \log X_{(k),i}}.$$

For each k, m these estimators are asymptotically *unbiased*, i.e. $\mathbb{E}[\xi_i] \sim \theta$ when θ goes to ∞. Their variances are all about $1/km$. Here we give a fourth estimator, which is also asymptotically unbiased:

$$\hat{\xi} = \frac{km-1}{\sum_{i=1}^{m} X_{(k),i}}.$$

Using information and estimation theories, we first show that the estimator $\hat{\xi}$ is optimal under a simplified model, the so-called *independent model*. Then we discuss its actual optimality.

2 The best estimation under the *independent model*

2.1 The independent model

Recall that a real-valued random variable follows the Gamma law with parameters (k, λ) if

$$\mathbb{P}(X \in [t, t+dt[) = \tfrac{t^{k-1}}{\Gamma(k)} \lambda^k e^{-\lambda t} \mathbf{1}_{t \geq 0} dt.$$

When θ goes to infinity, the asymptotic behavior of the minimum $X_{(1)}$ of θ random uniform variables in $[0, 1]$ is well-known (see for example [Fel70]) :

$$\theta X_{(1)} \xrightarrow[\theta \to \infty]{\mathcal{L}} \gamma_1,$$

where γ_1 follows the Gamma$(1, 1)$ law. More generally, one can prove here the following convergence :

Proposition 1. *With notations of Algorithm 1,*

$$(\theta X_{(k),1}, \ldots, \theta X_{(k),m}) \xrightarrow[\theta \to \infty]{\mathcal{L}} (\gamma_1, \ldots, \gamma_m), \tag{1}$$

where the γ_i are i.i.d. r.v. of law Gamma$(k, 1)$.

Proof. Let A be the event

$$A = A_{k,m,\theta} = \{ \text{ for all } i = 1, \dots, m, \text{ at least } k \text{ values fell in the } i\text{-th interval} \}.$$

Obviously, A occurs with high probability when θ is large. Indeed, it follows from a classical inequality (see [Bol85], Chap.4) involving Binomial r.v. that

$$\mathbf{P}(A) \geq 1 - 2me^{-\frac{\theta}{2m^2}}.$$

The main argument is that conditionned to A, the m-uplet $(X_{(k),1}, \dots, X_{(k),m})$ has a density. In order to lighten the notations, we will now write the prove with $m = 2$. Fix two real numbers μ and ν,

$$\begin{aligned}\mathbb{E}[e^{i\mu\theta X_{(k),1} + i\nu\theta X_{(k),2}}] &= \mathbb{E}[e^{i\mu\theta X_{(k),1} + i\nu\theta X_{(k),2}} \mathbf{1}_A] + \mathbb{E}[e^{i\mu\theta X_{(k),1} + i\nu\theta X_{(k),2}} (1 - \mathbf{1}_A)] \\ &= \mathbb{E}[e^{i\mu\theta X_{(k),1} + i\nu\theta X_{(k),2}} \mathbf{1}_A] + \mathrm{o}(1).\end{aligned}$$

$$\mathbb{E}[e^{i\mu\theta X_{(k),1} + i\nu\theta X_{(k),2}} \mathbf{1}_A] = \iint_{[0,\frac{1}{2}]^2} e^{i\mu\theta x + i\nu\theta y} \binom{\theta}{(k-1)(k-1)\ 1\ 1} x^{k-1} y^{k-1} dxdy(1-x-y)^{\theta-2k}.$$

Set $u = \theta x$ and $v = \theta y$, and finally use Stirling's formula :

$$\begin{aligned}&= \iint_{[0,\frac{1}{2}]^2} e^{i\mu u + i\nu v} \frac{\theta!}{(\theta - 2k)!} \left(\frac{u^{k-1}}{\theta^{k-1}\Gamma(k)} \right) \left(\frac{v^{k-1}}{\theta^{k-1}\Gamma(k)} \right) \frac{du}{\theta} \frac{dv}{\theta} \left(1 - \tfrac{u}{\theta} - \tfrac{v}{\theta}\right)^{\theta-2k} \\ &\to \iint_{[0,+\infty]^2} e^{i\mu u + i\nu v} \left(\frac{u^{k-1}}{\Gamma(k)} \right) \left(\frac{v^{k-1}}{\Gamma(k)} \right) e^{-2k-u-v} dudv \\ &= \mathbb{E}[e^{i\mu\gamma_1 + i\nu\gamma_2}],\end{aligned}$$

where γ_1 and γ_2 are i.i.d. Gamma$(k, 1)$ r.v; the convergence in Law is proved. □

We first assume that θ is large enough, and thus we will replace the actual $X_{(k),i}$ by their asymptotic limits. Since

$$\frac{1}{\theta} \text{ Gamma}(k, 1) \overset{\text{Law}}{=} \text{ Gamma}(k, \theta)$$

we assume in this section that the $X_{(k),i}$ are i.i.d. r.v. of law Gamma(k,θ), this is the so-called *independent model*. This is of course a strong assumption, because $X_{(k),i}$'s are not even independent.

2.2 Optimality

Proposition 2. *Under the independent model, $\hat{\xi}$ is unbiased.*

$$\mathbb{E}[\hat{\xi}] = \theta,$$
$$Var(\hat{\xi}) = \frac{\theta^2}{km-2}.$$

This is indeed better than the 3 estimators proposed in [Gir05].

Recall a few definitions in Statistics (see for example [Leh83]): given a m-sample of i.i.d. random variables $X_1, \dots, X_m$ of some law P_θ, any random variable $S = S(X_1, \dots, X_m)$ is called a *statistic*. Here the m-sample is $(X_{(k),1}, \dots, X_{(k),m})$ and $S = \sum_{i=1}^m X_{(k),i}$.

Definition 1. *A statistic S is* sufficient *for the parameter θ if, conditionnally to S, the law of $(X_1, \dots, X_m)$ does not depend on θ.*

More informally, S is sufficient if, given S, the knowledge of $(X_1, \dots, X_m)$ does not give any information on θ. Fortunately, there exists a simple criterion to check the sufficientness of a statistic, which may also be seen as a definition :

Proposition 3 (Neyman-Fisher). *Let $P_\theta(x_1, \dots, x_m)$ be the law of the m-sample. Assume that P_θ is absolutely continuous w.r.t. a measure μ. A statistic S is sufficient for the parameter θ if one may write*

$$dP_\theta = g(S(x_1, \dots, x_m), \theta) h(x_1, \dots, x_m) d\mu,$$

where g, h are non-negative measurable functions, h does not depend on θ.

Here, this criterion holds since

$$\mathbf{P}(X_1 \in [t_1, t_1 + dt_1), \dots X_m \in [t_m, t_m + dt_m)) = \prod_{i=1}^m \frac{{t_i}^{k-1}}{\Gamma(k)} \mathbf{1}_{t_1, \dots, t_m \geq 0} \; \theta^{mk} \; e^{-\theta S(t_1, \dots, t_m)} dt_1 \dots dt_m.$$

Definition 2. *A statistic S is* complete *if whenever $h(S)$ is a continuous non-negative function of S for which $\mathbb{E}[h(S)] = 0$ for all θ, then $h \equiv 0$, P_θ almost everywhere.*

It is easy to check that the statistic $S = \sum_{i=1}^m X_{(k),i}$ is complete. Completeness and sufficientness ensure optimality of the estimation :

Proposition 4 (Lehmann-Scheffé's Theorem). *Let S be a sufficient and complete statistic. Let ξ^* be another unbiased (i.e. $\mathbb{E}[\tilde{\xi}] = \theta$) estimator of θ. Among all the unbiased estimators of θ, $\mathbb{E}[\xi^*|S]$ has a minimal variance. Such an estimator is said to be* efficient.

Corollary 1. *Let $\tilde{\xi}$ another unbiased estimator of θ. Under the independent model,*

$$\mathbb{E}[(\tilde{\xi} - \theta)^2] \geq \mathbb{E}[(\hat{\xi} - \theta)^2].$$

This is a consequence of the Theorem, with $S = \sum_{i=1}^{m} X_{(k),i}$ and $\xi^* = \hat{\xi}$, because

$$\mathbb{E}[\hat{\xi}|S] = \hat{\xi}.$$

We have shown that under the independent model, $\hat{\xi}$ is optimal in the sense that among all the unbiased estimators, it has the lowest variance.

3 Optimality in the exact model

Let us return to the *exact model*: $X_{(p),i}$ is the p-th smallest realization of θ i.i.d. r.v. uniform on $[0, 1]$, among the values lying in $[\frac{i-1}{m}, \frac{i}{m}]$ shifted by $(i-1)/m$. Here is our result:

Theorem 1 (Optimality in the exact model). *Let $\tilde{\xi}(\mathbf{x})$, with $\mathbf{x} = (x_1, \cdots, x_m)$ another estimator of θ. Let $b(\theta)$ be the bias $\mathbb{E}_\theta[\tilde{\xi} - \theta]$. We assume that*

1. *$b(\theta) = \mathrm{O}(\sqrt{\theta})$.*
2. *$\tilde{\xi} : \mathbb{R}^m \to \mathbb{R}$ is continuous.*

Then

$$\mathbb{E}[(\tilde{\xi} - \theta)^2] \geq \mathbb{E}[(\hat{\xi} - \theta)^2] + \mathrm{O}(\theta).$$

Proof. We denote by $\mathbb{E}$ and $\mathbb{E}_{\text{ind}}$ the expectations corresponding respectively to the exact and independent models.

For any real $K > 0$, set $\tilde{\xi}_K = \tilde{\xi} \wedge K$.

$$\begin{aligned}
\mathbb{E}[(\tilde{\xi} - \theta)^2] &= \mathbb{E}[(\tilde{\xi} - \theta)^2 - (\tilde{\xi}_K - \theta)^2] && (2)\\
&+ \mathbb{E}[(\tilde{\xi}_K - \theta)^2] - \mathbb{E}_{\text{ind}}[(\tilde{\xi}_K - \theta)^2] && (3)\\
&+ \mathbb{E}_{\text{ind}}[(\tilde{\xi}_K - \theta)^2]. && (4)
\end{aligned}$$

Line (2). We may assume that $\tilde{\xi}^2$ is integrable, if not there is nothing to prove. One may write

$$\begin{aligned}(2) &= \mathbb{E}[\tilde{\xi}^2 - \tilde{\xi}_K^2] + 2\theta\mathbb{E}[\tilde{\xi}_K - \tilde{\xi}] \\ &\geq 2\theta\mathbb{E}[\tilde{\xi}_K - \tilde{\xi}].\end{aligned}$$

Then, if K is large enough (not depending on θ), $|(2)| \leq 2\theta$.

Line (3). Fix such a K, since $\tilde{\xi}_K$ is a bounded and continuous function, and using Proposition 1, $(3) = \mathrm{o}(1)$.

Line (4).

$$\begin{aligned}\mathbb{E}_{\mathrm{ind}}[(\tilde{\xi}_K - \theta)^2]\mathbb{E}_{\mathrm{ind}}[(\tilde{\xi}_K - \tilde{\xi})^2] + \mathbb{E}_{\mathrm{ind}}[(\tilde{\xi} - \theta)^2] \\ \geq \mathbb{E}_{\mathrm{ind}}[(\tilde{\xi} - \theta)^2] \\ \geq \mathbb{E}_{\mathrm{ind}}[(\hat{\xi} - \theta)^2].\end{aligned}$$

using Lehmann-Scheffé's Theorem. □

References

[Bol85] B. Bollobás. *Random graphs.* Academic Press, 1985.

[BYJK+02] Z. Bar-Yossef, T. Jayram, R. Kumar, D. Sivakumar, and L. Trevisan. Counting distinct elements in a data stream, 2002.

[DF93] Marianne Durand and Philippe Flajolet. Loglog counting of large cardinalities. *11th Annual European Symposium on Algorithms (ESA03)*, 1993.

[Fel70] William Feller. *An Introduction to Probability Theory and its Applications, Vol. I.* John Wiley & sons, 1970.

[Fla04] Philippe Flajolet. Counting by coin tossings. *ASIAN'04*, pages 1–12, 2004.

[FM85] Philippe Flajolet and Nigel Martin. Probabilistic counting algorithms for data base applications. *Journal of Computer and System Sciences, vol 31(2)*, pages 182–209, 1985.

[Gir05] Frédéric Giroire. Order statistics and estimating cardinalities of massive data sets. *DMTCS proceedings*, International Conference on Analysis of Algorithms:157–166, 2005.

[Knu73] Donald Knuth. *The Art of Computer Programming, vol. 3 : Sorting and Searching*. Addison-Wesley, 1973.

[Leh83] E.L. Lehmann. *Theory of Point Estimation*. John Wiley & sons, 1983.

[PD03] P.Indyk and D.Woodruff. Tight lower bounds for the distinct elements problem. *Proceedings of the 44th IEEE Symposium on Foundations of Computer Science, Boston*, pages 283–290, 2003.

Institut Élie Cartan Nancy (IECN)
Université Henri Poincaré Nancy I
BP 239 - 54506 Vandoeuvre-Les-Nancy Cedex - France

{chassain,gerin}@iecn.u-nancy.fr.