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

The aim of this note is to improve a solution proposed by Giroire [Gir05] to the following problem: consider a sequence $Y = (Y_1, \dots, Y_N)$ 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 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]). 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, 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 $\mathbf{X} = \{X_1, \dots, X_N\}$ 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 $\{Y_1, \dots, Y_N\}$ is equivalent to that contained in $(X_{(1)}, \dots, X_{(\theta)})$.

As a consequence, we are now dealing with a classical statistical problem: given a (small) sample of $(X_1, \dots, X_\theta)$, i.i.d. r.v., uniform on $[0,1]$, we want to estimate the (large) parameter θ. Denote by M the memory available (how many real numbers that can be stored). 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 $X_1, \dots, X_\theta$. 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] Version of February 2, 2008.

Giroire's 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, $1/X_{(2)}, 1/X_{(3)} \ldots$ 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.

Algorithm 1.
let k, m be integers. initialize $(X_{(1),i}, \ldots, X_{(k),i}, i = 1, \ldots, m)$ with $X_{(p),i} = \frac{i}{m}$ for all i, p.
for $j = 1$ to N
    $X_j = h(Y_j)$.
    let i the integer such that X_j lies in $[\frac{i-1}{m}, \frac{i}{m}[$.
    update the k-dimensional vector of k smallest values $X_{(1),i}, \ldots, X_{(k),i}$ lying in $[\frac{i-1}{m}, \frac{i}{m}[$.
next j.
for all p, i, renormalize $X_{(p),i} = m\big(X_{(p),i} - \frac{i-1}{m}\big)$.
return an estimator $\hat{\xi} = \hat{\xi}(X_{(l),i}; i = 1, \ldots, m; l = 1, \ldots, k)$.

Thus we get m vectors in $\mathbb{R}^k$. $X_{(k),i}$ is the k-th smallest hashed value lying in $[\frac{i-1}{m}, \frac{i}{m}]$, renormalized to get a real in $[0,1]$. If less than l values have fell 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 function, square root function 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:

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

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

2 The best estimation under the *independant model*

In this section, $\Rightarrow$ denotes the convergence in law. Recall that a real-valued random variable is said to follow the Gamma law with parameters (k, θ) if

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

The asymptotic behavior of the minimum $X_{(1)}$ of θ random uniform variables in $[0,1]$ is well-known (see for example [Fel70]): $X_{(1)} \Rightarrow \gamma_1$, where γ_1 follows the Gamma$(1, \theta)$ law. More generally, one can prove here the following convergence

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

where the γ_i are i.i.d. r.v. of law Gamma(k,1). Consequently, 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*. We set

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

Remark 1. *This estimator depends only on the m values $(X_{(k),i}, i = 1, \dots, m)$, not on the $km - m$ other hashed values stored by the algorithm. This follows from the fact that the knowledge of these values does not provide any information about θ. For a given i, conditionnally on $X_{(k),i}$, the r.v. $(X_{(1),i}, \dots, X_{(k-1),i})$ are distributed uniformally on $[0, X_{(k),i}]$.*

A simple calculation shows that under the independent model,

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

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 we consider the statistic $S = \sum_{i=1}^m X_{(k),i}$.

Definition 1. *A statistic S is* sufficient *for the parameter θ if and only 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. $\hat{S}$ is sufficient.

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

Here, the statistic $S = \sum_{i=1}^m X_{(k),i}$ is complete and sufficient. Some simple criterions to check sufficientness and completeness are given in [Leh83]. Complete sufficient statistics share the following useful property:

Theorem 1 (Lehmann-Scheffé)**.** *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,*

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

Remark 2. *Note that $Var(\hat{\xi})$ is about θ^2. This optimal bound does not depend on the algorithm, see [PD03].*

3 Optimality in the real model

From now one places oneself in 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 $[\frac{i-1}{m}, \frac{i}{m}]$. When $i \neq j$, there is now dependancy between $X_{(k),i}$ and $X_{(k),j}$. Set $\mathbf{P}$ and $\mathbf{P}_{\text{ind}}$ the laws corresponding respectively to the exact and independant models, $\mathbb{E}$ and $\mathbb{E}_{\text{ind}}$ the corresponding expectations.

Lemma 1. *Let A be the event*

$$A = A_{k,m,\theta} : \text{“for all } i = 1, \dots, m,\ X_{(k),i} < \tfrac{1}{m}\text{”}.$$

(i.e. *at least k hashed values have falled in each of the m intervals). From a classical inequality (see [Bol85]) we get*

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

Here is the main result:

Theorem 2 (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 $\mathbb{E}_\theta[\tilde{\theta} - \theta]$. We assume*

1. *$b(\theta) = \mathrm{O}(\sqrt{\theta})$.*
2. *There exists a constant C such that $|\tilde{\xi}(\mathbf{x})| \leq \frac{C}{\|\mathbf{x}\|}$, for every $\mathbf{x}$ in $\mathbb{R}^m$, $\| \mathbf{x} \|$ large enough.*

Then

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

Proof. First write

$$\begin{aligned}\mathbb{E}[(\tilde\theta-\theta)^2] &= \mathbb{E}[(\tilde\theta-\theta)^2\mathbf{1}_A] + \mathbb{E}[(\tilde\theta-\theta)^2(1-\mathbf{1}_A)],\\ &= \mathbb{E}[(\tilde\theta-\theta)^2\mathbf{1}_A] + o(\theta),\end{aligned}$$

using Lemma 1. We now bring back ourselves to the independant model:

$$\mathbb{E}[(\tilde\theta-\theta)^2] = \mathbb{E}_{\text{indé}}[(\tilde\theta-\theta)^2] + \big(\mathbb{E}[(\tilde\theta-\theta)^2\mathbf{1}_A] - \mathbb{E}_{\text{indé}}[(\tilde\theta-\theta)^2]\big) + o(A), \tag{2}$$

The key point is the fact that conditioned to the event A, the r.v. $(X_{(k),1},\dots,X_{(k),m})$ admit a density toward the Lebesgue measure on $\mathbb{R}_+$.

$$\begin{aligned}&\mathbb{E}[(\tilde\theta-\theta)^2] - \mathbb{E}_{\text{indé}}[(\tilde\theta-\theta)^2] = \\ &\quad\int_{[0,\frac{1}{m}]^m}\binom{\theta}{(k-1)\dots(k-1)\ \ 1\dots1}{x_1}^{k-1}\dots{x_m}^{k-1}(1-x_1-\dots-x_m)^{\theta-mk}dx_1\dots dx_m\\ &\qquad - \int_{\mathbb{R}_+^m}\tfrac{{x_1}^{k-1}}{\Gamma(k)}\dots\tfrac{{x_m}^{k-1}}{\Gamma(k)}\theta^{km}e^{-\theta(x_1+\dots+x_m)}dx_1\dots dx_m.\end{aligned}$$

We omit the proof of the following lemma:

Lemma 2.

$$\left|\tfrac{\theta(\theta-1)\cdots(\theta-2k+1)}{\theta^{2k}}(1-x_1-\dots-x_m)^{\theta-2k} - e^{-\theta(x_1+\dots+x_m)}\right| \leq c^{ste}\theta(x_1+\dots+x_m)^2e^{-\theta(x_1+\dots+x_m)},$$

where the constant depends neither on θ nor on the x_i's.

Hence

$$\begin{aligned}&|\mathbb{E}[(\tilde\theta-\theta)^2] - \mathbb{E}_{\text{indé}}[(\tilde\theta-\theta)^2]| \leq \\ &\int_{[0,\frac{1}{m}]^m}|\tilde\theta(x_1,\dots,x_m)-\theta|^2\tfrac{{x_1}^{k-1}}{\Gamma(k)}\dots\tfrac{{x_m}^{k-1}}{\Gamma(k)}\theta^{km}\left\{c^{\text{ste}}\theta(x_1+\dots+x_m)^2e^{-\theta(x_1+\dots+x_m)}\right\}dx_1\dots dx_m + \mathrm{O}(\theta)\end{aligned}$$

Set $y_i=\theta x_i, i=1\dots m$ in the integrand:

$$\leq \int_{[0,\frac{\theta}{m}]^m}|\tilde\theta(\tfrac{y_1}{\theta},\dots,\tfrac{y_m}{\theta})-\theta|^2\tfrac{{x_1}^{k-1}}{\Gamma(k)}\dots\tfrac{{x_m}^{k-1}}{\Gamma(k)}\theta^{-1}\left\{c^{\text{ste}}\theta(y_1+\dots+y_m)^2e^{-\theta(y_1+\dots+y_m)}\right\}dy_1\dots dy_m + \mathrm{O}(\theta)$$

Here we use the hypothesis made on the estimator: $\theta(\mathbf{x}) \leq \frac{1}{\|\mathbf{x}\|}$. We also need the following arithmetico-geometric inequality:

$$(a_1\dots a_m)^\alpha \leq \lambda({a_1}^2+\dots+{a_m}^2)^{m\alpha/2}.$$

Set $\mathbf{y}=(y_1,\dots,y_m)$, one gets

$$\begin{aligned}&\leq c^{\text{ste}}\int_{[0,\frac{\theta}{m}]^m}\frac{\theta^2}{\|\mathbf{y}\|^2}(\|\mathbf{y}\|)^{m(k-1)}\theta^{-1}(y_1+\dots+y_m)^2e^{-y_1-\dots-y_m}dy_1\dots dy_m + \mathrm{O}(\theta)\\ &\leq c^{\text{ste}}\int_{[0,\frac{\theta}{m}]^m}\frac{\theta^2}{\|\mathbf{y}\|^2}(\|\mathbf{y}\|)^{m(k-1)}\theta^{-1}(\|\mathbf{y}\|)^2e^{-\|\mathbf{y}\|}dy_1\dots dy_m + \mathrm{O}(\theta).\end{aligned}$$

Here we make a "polar-like" change of variables in $\mathbb{R}^m$. We get this inequality:

$$\begin{aligned}|\mathbb{E}[(\tilde\theta-\theta)^2] - \mathbb{E}_{\text{indé}}[(\tilde\theta-\theta)^2]| &\leq c^{\text{ste}}\theta\int_0^{\sqrt{m}\frac{\theta}{m}}r^\alpha e^{-r}dr + \mathrm{O}(\theta), \quad \alpha>0.\\ &= \mathrm{O}(\theta).\end{aligned}$$

(2) has become

$$\begin{aligned}\mathbb{E}[(\tilde{\theta}-\theta)^2] &= \mathbb{E}_{\text{indé}}[(\tilde{\theta}-\theta)^2] + \mathrm{O}(\theta)\\ &= \mathbb{E}_{\text{indé}}[(\hat{\theta}-b(\theta)-\theta)^2] + b^2(\theta) + \mathrm{O}(\theta)\\ &\geq \mathbb{E}_{\text{indé}}[(\hat{\theta}-\theta)^2] + \mathrm{O}(\theta),\end{aligned}$$

□

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.