Efficient estimation of the cardinality of large data sets

Philippe Chassaing, Lucas Gerin

Fourth Colloquium on Mathematics and Computer Science DMTCS proc. **AG**, 2006, 419–422

Efficient estimation of the cardinality of large data sets

Philippe Chassaing[1] and Lucas Gerin[1]

[1] *Institut Élie Cartan Nancy,Université Henri Poincaré Nancy 1. B.P. 239, F-54506 Vandoeuvre-lès-Nancy Cedex, France*

Giroire (Gir05) 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 $[0,1]$. In this note we propose an optimal estimation, using Kullback information and estimation theory.

Keywords: cardinality,large multiset,approximate counting

1 Introduction

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 (often 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)). Usual 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. A possible 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.

Probabilistic Counting 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 $\mathbf{X} = \{X_1, \ldots, X_N\}$ 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 (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 the 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).

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, $1/X_{(2)}, 1/X_{(3)} \dots$ 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.

Algorithm 1
let k, m be integers. initialize $(X_{(1),i}, \dots, X_{(k),i}, i = 1, \dots, m)$ 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 $X_{(1),i}, \dots, X_{(k),i}$ 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 $\hat{\xi} = \hat{\xi}(X_{(l),i}; i = 1, \dots, m; l = 1, \dots, k)$.

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 $X_1, \dots, X_N$.

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*

Recall that a real-valued random variable X 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)): $\theta X_{(1)} \xrightarrow{\mathcal{L}} \gamma_1$, where γ_1 follows the Gamma$(1, \theta)$ law. More generally, we can prove here the following convergence:

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

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 2.1 This estimator depends only on the m values $(X_{(k),i}, i = 1, \dots, m)$, not on the $m(k-1)$ other hashed values stored by the algorithm. This follows from the fact that the knowledge of these values does not provide additional information on θ: for a given i, conditionnally on $X_{(k),i}$, the r.v. $(\frac{X_{(1),i}}{X_{(k),i}}, \dots, \frac{X_{(k-1),i}}{X_{(k),i}})$ are distributed uniformly on $[0,1]$.

A simple calculation shows that under the independent model,

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

This is indeed better than the 3 estimators proposed in (Gir05). We can now use the powerful information theory. One calls *statistic* any random variable which is a function of the sample (here, $S = \sum_{i=1}^{m} X_{(k),i}$ is a statistic).

Theorem 1 (Lehmann-Scheffé) *Let S be a sufficient and complete statistic. Let ξ^* be an unbiased estimator of θ (i.e. $\mathbb{E}[\xi^*] = \theta$). Among all the unbiased estimators of θ, $\mathbb{E}[\xi^*|S]$ has a minimal variance. Such an estimator is said to be* efficient.

For the definitions of sufficientness and completeness, one may read for example (Leh83). Here S is sufficient and complete.

Corollary 1 (Optimality in the independent model) *Let $\tilde{\xi}$ be another unbiased estimator of θ. Under the independant model,*

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

3 Optimality in the real model

From now on, we consider the *real 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 $[\frac{i-1}{m}, \frac{i}{m}]$. For all i, j, there is now dependency between $X_{(k),i}$ and $X_{(k),j}$. We can no more use directly information theory.

Theorem 2 (Optimality in the exact model) *Set $\hat{\xi} = \frac{km-1}{\sum_{i=1}^{m} X_{(k),i}}$. Let $\tilde{\xi}(\mathbf{X})$, with $\mathbf{X} = (X_{(k),1}, \ldots, X_{(k),m})$, be another estimator of θ. We assume that $b(\theta) := \mathbb{E}_\theta[\tilde{\xi} - \theta] = \mathrm{O}(\sqrt{\theta})$. Then*

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

Remark 3.1 We have already seen that $\text{Var}(\hat{\xi})$ is about θ^2. It has been shown (IW03) that it is an optimal universal bound.

Proof: There are mainly two steps in the proof:

1. It can be shown that then θ is large, the "good" case (i.e. at least k real fall in each interval) occurs with high probability.
2. The convergence in (1) has then to be studied in details.

□

References

[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.

[IW03] 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.

[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.

[NAS96] Y. Matias N. Alon and M. Szegedy. The space complexity of approximating the frequency moments. *Proceedings of the 28th ACM Symp. on Theory of Computing*, pages 20–29, 1996.