Adaptive sampling

Adaptive sampling [a1] is a probabilistic algorithm invented by M. Wegman (unpublished) around 1980. It provides an unbiased estimator of the number of distinct elements (the "cardinality") of a file (a sequence of data items) of potentially large size that contains unpredictable replications. The algorithm is useful in data-base query optimization and in information retrieval. By standard hashing techniques [a3], [a6] the problem reduces to the following.

A sequence $(h_1, \ldots, h_n)$ of real numbers is given. The sequence has been formed by drawing independently and randomly an unknown number N of real numbers from $[0, 1]$, after which the elements are replicated and permuted in some unknown fashion. The problem is to estimate the cardinality N in a computationally efficient manner.

Three algorithms can perform this task.

1) Straight scan computes incrementally the sets $U_j = \{h_1, \ldots, h_j\}$, where replications are eliminated on the fly. (This can be achieved by keeping the successive U_j in sorted order.) The cardinality is then determined exactly by $N = \operatorname{card}(U_n)$ but the auxiliary memory needed is N, which may be as large as n, resulting in a complexity that is prohibitive in many applications.

2) Static sampling is based on a fixed sampling ratio p, where $0 < p \leq 1$ (e.g., $p = 1/100$). One computes sequentially the samples $U_j^* = \{h_1, \ldots, h_j\} \cap [0, p]$. The cardinality estimate returned is $N^* = \operatorname{card}(U_n^*)/p$. The estimator N^* is unbiased and the memory used is Np on average.

3) Adaptive sampling is based on a design parameter $b \geq 2$ (e.g., $b = 100$) and it maintains a dynamically changing sampling rate p and a sequence of samples U_j^{**}. Initially, $p = 1$ and $U_0^{**} = \emptyset$. The rule is like that of static sampling, but with p divided by 2 each time the cardinality of U_j^{**} would exceed b and with U_j^{**} modified accordingly in order to contain only $U_j \cap [0, p]$. The estimator $N^{**} = \operatorname{card}(U_n^{**})/p$ (where the final value of p is used) is proved to be unbiased and the memory used is at most b.

The accuracy of any such unbiased estimator $\tilde{N}$ of N is measured by the standard deviation of $\tilde{N}$ divided by N. For adaptive sampling, the accuracy is almost constant as a function of N and asymptotically close to

$$\frac{1.20}{\sqrt{b}},$$

a result established in [a1] by generating functions and Mellin transform techniques. An alternative algorithm, called probabilistic counting [a2], provides an estimator N^{***} of cardinalities that is unbiased only asymptotically but has a better accuracy, of about $0.78/\sqrt{b}$.

Typically, the adaptive sampling algorithm can be applied to gather statistics on word usage in a large text. Well-designed hashing transformations are then known to fulfill practically the uniformity assumption [a4]. A general perspective on probabilistic algorithms may be found in [a5].

References

[a1] P. Flajolet, "On adaptive sampling" *Computing* , **34** (1990) pp. 391–400

[a2] P. Flajolet, G.N. Martin, "Probabilistic counting algorithms for data base applications" *J. Comp. System Sci.* , **31** : 2 (1985) pp. 182–209

[a3] D.E. Knuth, "The art of computer programming" , **3. Sorting and Searching** , Addison-Wesley (1973)

[a4] V.Y. Lum, P.S.T. Yuen, M. Dodd, "Key to address transformations: a fundamental study based on large existing format files" *Commun. ACM* , **14** (1971) pp. 228–239

[a5] R. Motwani, P. Raghavan, "Randomized algorithms" , Cambridge Univ. Press (1995)

[a6] R. Sedgewick, "Algorithms" , Addison-Wesley (1988) (Edition: Second)

Ph. Flajolet

This text originally appeared in Encyclopaedia of Mathematics - ISBN 1402006098

Copyright © 2001 All rights reserved. Privacy Policy | Terms of use