A Supplement to Sampling-Based Methods for Query Size Estimation in a Database System

Yibei Ling and Wei Sun
School of Computer Science
Florida International University
Miami, Florida 33199, USA.
Internet: {weisun,lingy}@fiu.edu

abstract

Sampling-based methods for estimating relation sizes after relational operators such as selections, joins and projections have been intensively studied in recent years. Methods of this type can achieve high estimation accuracy and efficiency. Since the dominating overhead involved in a sampling-based method is the sampling cost, different variants of sampling methods are proposed so as to minimize the sampling percentage (thus reducing the sampling cost) while maintaining the estimation accuracy in terms of the confidence level *and* relative error *(to be precisely defined later in* Section 2*). In order to determine the minimal sampling percentage, the overall characteristics of the data such as the mean and variance are needed. Currently, the representative sampling-based methods in literature are based on the assumption that overall characteristics of data are unavailable, and thus a significant amount of effort is dedicated to estimating these characteristics so as to approach the optimal (minimal) sampling percentage. The estimation for these characteristics incurs cost as well as suffers the estimation error. In this short essay, we point out that the exact values of these characteristics of data can be kept track of in a database system at a negligible overhead. As a result, the minimal sampling percentage while ensuring the specified relative error and confidence level can be precisely determined.*

1 Introduction

The central task of sampling-based methods is to estimate the size of the (intermediate) relation (relations) after applying a relational operator such as a selection, projection and/or join to a relation (relations). We call this task *(query) size estimation*. Accurate size estimation is essential to a query optimizer in a database system. In addition, size estimation is also needed for answering certain statistics retrievals in a database system.

There are different estimation methods: *parametric methods* [1, 2, 12], *table-based models* [13, 3], and *sampling-based methods* [7, 8, 5, 9, 10, 11]. Unlike table-based methods, the sampling-based methods do not need to extract, store or maintain the summary information about data in a database. In many cases, extracting, storing and maintaining these summary data is rather costly. Unlike parametric methods, the sampling-based methods do not make a priori assumption about statistical characteristic of data such as the distribution and/or correlation. As a result, a significantly more accurate size estimation can be achieved.

Sampling-based methods have been popular in recent years because of their estimation accuracy. However, for sampling-based methods, achieving good estimation accuracy and high efficiency (of estimation) is inherently conflicting: for lowering down sampling overhead, a smaller sampling percentage is sought, thus a less accurate estimation is made; for achieving a good estimation accuracy, the representativeness of the whole population by the samples must be ensured, which in turn requires a larger sampling percentage. Basically, there are three representative sampling-based approaches in the literature.

Adaptive sampling: This is proposed by Lipton, Naughton, and Schneider [11, 10]. This method is intended to meet the desired estimation accuracy

in pursuit of minimal sampling cost. The termination conditions for sampling are determined by the overall statistical characteristics of data (precisely the means and variances), the confidence level, the amount of the samples, and the characteristics of the samples. Thus, the sampling percentage is *dynamically* or *adaptively* determined (while the sampling is in progress), instead of a constant percentage number. Since the characteristics of the population are usually not priori known, some approximating values (say, by upper bounds) are used for these parameters, which in turn may lead to a greater sampling overhead.

Double sampling or two-phase sampling: W. Hou, G. Ozsoyoglu and E. Dogdu [6] makes use of the *double sampling* [4, 14] technique which has been well-defined in statistics in order to reduce the sampling cost while retaining the estimation accuracy specified. The sampling is conducted in two stages. The first stage of sampling is to obtain (estimate) preliminary information of the data population such as the mean and variance. Based on the information obtained in the first stage, the amount of sampling needed for the second stage sampling is determined. Samples in the second stage are used for size estimation such that the estimated results are within the specified error bound and within a given confidence level. A constant sampling percentage of 2% for the first stage is used. However, no theoretical guideline are provided to determine the amount of sampling in the first stage. Clearly, a good estimation of the statistical characteristics of the data in the first stage is essential to the estimation accuracy.

Sequential sampling: This is proposed by P. Haas and A. Swami [5] which improves on *double sampling* and the *adaptive sampling* by estimating the overall characteristics of the population (the total mean and variance) using all tuples that have been sampled so far. This estimation is conducted in concurrency with the sampling. Again, the termination condition of the approach is determined at every sampling step by using *stopped random walks* [5]. Therefore, sampling will be terminated as long as the desired accuracy and confidence level is reached. The approach has been shown to be asymptotically efficient [5], and is superior to the other sampling methodologies in the literature. However, the computational overhead is incurred at every time a tuple is sampled. In addition, since the parameters of the population are estimated by using samples, certain degree of estimation errors are inevitable.

It can be observed that all the above methods are based on the assumption that we have not known a priori the overall statistical characteristic of the population such as the mean and variance. Therefore, a tremendous effort has been dedicated to the estimation of them using various methods. In this short essay, we show that exact values of these parameters can be kept track of with negligible effort and overhead. As a result, the lower bound of sampling percentage while retaining a given estimation accuracy and confidence level specified can be achieved. These proposed sampling-based methods seem to have directly adopted pure statistical sampling methods where normally no priori characteristics are available. Keeping tack of other similar parameter values such as the cardinality of a relation and the number of distinct values under an attribute for query processing and optimisation is standard in a database system, and the collection of them is normally called a database *profile*.

2 A Brief Introduction to Sampling-Based Methods

In this section, we shall briefly discuss the theoretical background of the sampling methods. Consider a population (namely all the values under an attribute of a relation instance), its statistical characteristics of population (i.e., the variance S and the mean $\overline{Y}$) can be computed from the given population. Let $\{y_1, y_2, ..., y_i, ..., y_N\}$ be all the individual values in the population, where N is the *population size* (the number of tuples in the relation instance). Let m be the *sample size* (the number of tuples sampled), $Z = \{z_1, z_2, ..., z_m\}$ be the sampled tuples such that $z_i, 1 \leq i \leq m$ is randomly chosen from the population with replacement, $Y = \sum_{i=1}^{N} y_i$ the *population total*, $\overline{Y} = \frac{\sum_{i=1}^{N} y_i}{N}$ the *population mean*, $S = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \overline{Y})^2}{N-1}}$ the *standard deviation* of the population, $\overline{z} = \frac{\sum_{i=1}^{m} z_i}{m}$ the *sample mean*, and $\frac{\overline{Y} - \overline{z}}{\overline{Y}}$ the *relative error* for the estimation. It is direct that the relative error of an estimation measures how accurate the estimation is.

The probability α that the relative error is greater than some given constant r (i.e., the confidence level) can be expressed as follows:

$$Pr(|\frac{\overline{Y} - \frac{\sum_{i=1}^{m} z_i}{m}}{\overline{Y}}| \geq r) = Pr(|\frac{\overline{Y} - \overline{z}}{\overline{Y}}| \geq r) = \alpha$$

Equivalently,

$$Pr(|\overline{Y} - \overline{z}| \geq r\overline{Y}) = Pr(\frac{|\overline{Y} - \overline{z}|}{\sigma_{\overline{z}}} \geq \frac{r\overline{Y}}{\sigma_{\overline{z}}}) = \alpha$$

where $\sigma_{\overline{z}} = \sqrt{\frac{N-m}{N}}\frac{S}{\sqrt{m}}$ is called the *sample standard error*, and α is considered to be the *confidence level* of the estimation. Assume that $\overline{z}$ is normally distributed, then the following formula can be observed [14]:

$$r\overline{Y} = t_\alpha \sigma_{\overline{z}} = \sqrt{\frac{N-m}{N}}\frac{S}{\sqrt{m}}$$

where $t_\alpha = \Phi^{-1}((1+\alpha)/2)$, Φ is the standardized normal random variable. Solving the above formula for m gives

$$m = (\frac{t_\alpha S}{r\overline{Y}})^2 / [1 + \frac{1}{N}(\frac{t_\alpha S}{r\overline{Y}})^2] \qquad (1)$$

The above computed m gives the minimal amount of sampling to be taken such that the confidence level α and the relative error r are ensured.

By the above formula, the sampling size m is determined by N, α, r, S and $\overline{Y}$. α and r are given by a user, and N is normally known for the given relation instance in the database profile. Thus, S and $\overline{Y}$ are needed in order to compute m.

The three representative sampling-based methods basically use different strategies to estimate the population mean $\overline{Y}$ and the variance S or provide certain bounds to them, so that an optimal (minimal) sampling percentage can be approached. Precisely, in the adaptive sampling method, its summation termination condition of sampling is $s > k_1(S/\overline{Y})d(d+1)$, where s is the sum of the samples, d and k_1 are associated with the relative error and the confidence level. Since certain characteristics of the population are unknown, certain upper bound b ($b > S/\overline{Y}$) are used to substitute $S/\overline{Y}$. Therefore, in certain cases the *adative sampling* may become extremely inefficient[5].

The *double sampling* is a classical method widely used in statistics [4, 14], and is an appropriate sampling method when the information about total statistical characteristics of population such as mean and variance are absent. It consists of two stages of sampling. The sampling in the first stage is to estimate the total mean and variance. Then, the $\overline{z}$ and S_z are used in the following formula in order to calculate the amount of sampling needed to meet the accuracy requirements.

$$m = \frac{(S_z \cdot t_\alpha)^2}{(r\overline{z})^2}\left(1 + 8\frac{r}{t_\alpha} + \frac{S_z^2}{m_1\overline{z}^2} + \frac{2}{m_1}\right) \qquad (2)$$

where $\overline{z} = \frac{\sum_{i=1}^{m_1} z_i}{m_1}$, $S_z = \sqrt{\frac{\sum_{i=1}^{m_1}(z_i - \overline{z})^2}{m_1}}$, and m_1 denotes the number of sampled tuples which are used to estimate the total mean and variance in the first step. We note that *Equation 2*, instead of *Equation 1*, is used to estimate m, because certain corrections should be made if S and $\overline{Y}$ are estimated values[14, 4]. If $m < m_1$, the sampling procedure terminates (implying that the sampling is enough), and thus estimation for the population can be accordingly made by the samples. if $m > m_1$, then additional amount of sampling $m - m_1$ in the second stage must be taken in order to meet the prespecified accuracy and confidence level. The problem in *double sampling* is that there is no theoretical guidance to determine the amount of sampling in the first stage. Hou, Ozsoyoglu and Dogdu [6] simply uses a 2% sampling percentage.

The *sequential sampling* by Haas and Swami [5] improves on the *double sampling* by estimating the total mean and variance at each sampling step and using all the samples taken so far. A guidline is established for the amount of sampling for an accurate estimation of the over characteristics of the population. Their experimental and theoretical results [5] demonstrate that this approach is more adaptive and more parsimony in terms of the amount of sampling.

3 Maintaining Exact Characteristics at Negligible Overhead

As we can see that the overall mean and variance of the population is necessary for a sampling-based method to determine the optimal sampling percentage under the specified confidence level and relative error. Unlike these methodologies in literature that use different strategies to estimate these characteristics, we point out in this short essay that the exact values of them can be kept track of with little effort even in the presence of deletion, insertion, and update of a tuple

or tuples in a database system. Collecting certain simpler characteristics such as the largest/smallest values and the number of distinct values under an attribute is popular in a database system.

Clearly, it is sufficient for us to focus on the deletion and insertion of a tuple, because an update of a tuple is equivalent to a deletion followed by an insertion, and the deletion and insertion of multiple tuples can be handled individually. The following formula can be directly observed. Therefore, detailed derivations of them are not presented.

For insertion of a tuple: Let y_{new} denote the value of the concerned attribute of the tuple to be inserted, $\overline{Y_{new}}$, $\overline{Y_{old}}$, S_{new} and S_{old} denote the means and variances of total population after and before the *Insert* operation. The following is obvious:

$$(1)\overline{Y_{new}} = \frac{N\overline{Y_{old}}+y_{new}}{N+1}$$

$$(2)S^2_{new} = \frac{N-1}{N}S^2_{old} + \frac{(\overline{Y_{old}}-y_{new})^2}{N+1}$$

$$(3)N = N+1$$

For deletion of a tuple: Let y_{old} denote the value of the concerned attribute of the tuple to be deleted, $\overline{Y_{new}}$, $\overline{Y_{old}}$, S_{new} and S_{old} denote the means and variances of total population after and before the *delete* operation. The following is obvious:

$$(1)\overline{Y_{new}} = \frac{N\overline{Y_{old}}-y_{old}}{N-1}$$

$$(2)S^2_{new} = \frac{N-1}{N-2}S^2_{old} - \frac{N\cdot(\overline{Y_{old}}-y_{old})^2}{(N-1)(N-2)}$$

$$(3)N = N-1$$

It can be easily observed that negligible computational and storage overhead is needed in keeping track of these characteristics. In most cases, these database system files are maintained in the main memory. Thus, essentially no page I/Os will be generated.

4 Conclusion

We point out that the the amount of sampling for query size estimation can be reduced to its minimum by keeping track of some summary information about data such as the mean and variance of data at a negligible overhead even in the presence of database updates. This will avoid unnecessary computation and sampling overheads of estimating the overall characteristics of data in the representative sampling-based methods.

References

[1] Stavros Christodoulakis. Estimating record selectivities. *Information Systems*, 8(2):105–115, 1983.

[2] Stavros Christodoulakis. Estimating block selectivities. *Information Systems*, 9(1):69–79, 1984.

[3] Pai-Cheng Chu. A contingency approach to estimating record selectivities. *Software Engineering*, 17(6):544–552, 1991.

[4] William G. Cochran. *Sampling Techniques*. John Wiley & Sons, 1977.

[5] Peter J. Haas and Arun N. Swami. Sequential sampling procedures for query size estimation. In *Proceedings of the Very Large Database Conference*, pages 341–350, April 1992.

[6] Wen-Chi Hou, G. Ossoyoglu, , and E. Dogdu. Error-constrained count query. evaluation in relational databases. In *Proceedings of the ACM-SIGMOD Conference*, pages 278–287, August 1991.

[7] Wen-Chi Hou and G. Ossoyoglu. Statistical estimators for aggregate relational algebra queries. *ACM Transactions On Database Systems*, 16(4):600–654, December 1991.

[8] Wen-Chi Hou, G. Ossoyoglu, and Baldeao K. Taneja. Processing aggregate relational queries with hard time constraints. In *Proceedings of the ACM-SIGMOD Conference*, pages 165–172, August 1989.

[9] Richard Lipton and Jefferey Naughton. Estimating the size of generalised transitive closures. In *Proceedings of the 15th VLDS Conference*, pages 165–172, 1989.

[10] Richard Lipton and Jefferey Naughton. Query size estimation by adaptive sampling. In *Proceedings of 9th ACM Symposium on Priciples of Database Systems*, pages 40–46, March 1990.

[11] Richard Lipton, Jeffery Naughton, and Donavan Schneider. Practical selectivity estimation through adaptive sampling. In *Proceedings of ACM SIGMOD*, pages 1–11, 1990.

[12] Clifford A. Lynch. Selectivity estimation and query optimisation in large databases with highly skewed distributions of column values. In *Proceedings of the 14th VLDS Conference*, pages 240–251, 1988.

[13] M. Muralikrishna and D. DeWitt. Statistical profile estimation in database system. *Computing Survey*, 20(3):191–221, 1988.

[14] P. V. Sukhatme and B. V. Sukhatme. *Sampling Theory of Surveys with Application*. Iowa State University Press, 1970.