On-Line Adaptive Sampling

Francesco Di Tria
Ministero dell'Istruzione
francesco.ditria1@posta.istruzione.it

Abstract. Sampling is a method for obtaining approximate answers using compressed and reduced data. Such a method may be used in decision making when fast answers are required or as a perturbation technique when data are to be obfuscated. In the paper, an adaptive sampling method is proposed. This solution is based on the analysis of the coefficient of variation that determines the sample size.

1 Introduction

Sampling is a well-known method that reduces the cardinality of a database and creates a concise representation of the original data, which is, then, used to perform statistical analyses, avoiding access to the original data. Since the sample is smaller than the source, this method is usually used for obtaining fast and approximate answers to aggregate queries [1].

An application field of sampling is the *decision making*, where the total precision is not the main aim, while fast answering times are highly required [4]. Another application field is the *security of statistical databases*. In this context, sampling is used as a perturbation technique in order to obfuscate individual data. Also in this case, the total accuracy is not required. On the other hand, it is mandatory avoiding statistical inference operations that can lead, through aggregate queries, to obtain data of a specific individual. So, approximate answers are a natural way to obtain obfuscated data, instead of adding noise after the query computation, which requires further elaboration time [3].

There are several methods for sampling and all of these aim at providing reliable approximate answers [2]. This means producing, along with the approximate answers, also an error bound. Usually, the error bound is a confidence interval computed on the basis of the variance. The contribution of the paper is a sampling method that does not use the variance *a posteriori* for estimating the confidence interval, but uses the variance *a priori* for determining an optimal sample size [6]. The rationale is that the greater is the variance, the greater should be the sample size in order to obtain a representative sample. So, the proposed method can be considered *adaptive*, in the sense that the sample size is not static but dynamically-computed on the basis of preliminary experimental data, that is the statistical profile of the database. Furthermore, differently from traditional methods that create a data synopsis from the source database, the proposed method has been implemented in an approximate query answering system that performs also an *on-line* sampling, in the sense that approximate answers are derived from an in-memory dataset. The system is publicly available at [7] as an open source project.

In what follows, first we introduce traditional methods and, then, we present our proposal.

2 Sampling methods and system architecture

The general architecture of a sampling-based system is depicted in Figure 1. First, a data reduction of a source database is executed. In this phase, a data synopsis is created using a sampling method. The data synopsis, whose size is several orders of magnitude lower than the source database, is, then, used to answer in a fast and an approximate way to statistical analyses.

The data synopsis is to be aligned with the data source and, therefore, is refreshed at regular intervals.

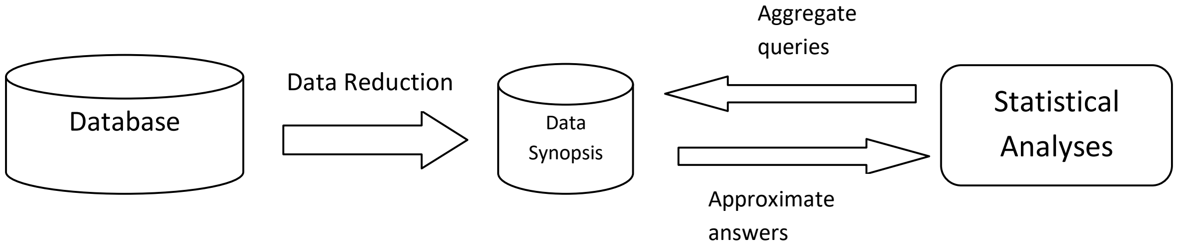

Figure 1. Sampling architecture.

Given a relation *R* of cardinality *n*, traditional sampling methods are:

- *Simple Random Sampling With Replacement*. A random number *r* between 1 and *n* is generated and, then, the *r*-th tuple of *R* is added to the sample. So, each tuple of *R* can appear more times in the sample. If each tuple has the same selection probability, then the sample is *unbiased*.
- *Simple Random Sampling Without Replacement*. A random number *r* between 1 and *n* is generated and the *r*-th tuple of *R* is added to the sample if and only if it has not previously added. So, each tuple of *R* can appear only once in the sample. To bias the sampling process, tuples are first grouped into subclasses and, then, a different number of tuples is chosen for each subclass. This variant is called *stratification*.
- *Poisson Sampling*. For each tuple of *R*, a random number *r* between 0 and 1 is generated. If *r* is lower than a threshold value, then the tuple is added to the sample. If the threshold is the same for all the tuples, then the sample is said *unbiased*, *biased* otherwise. In *Bernoulli Sampling*, the random values can be only 0 and 1, and this sampling method is equivalent to flipping a coin. Differently from other sampling methods, the sample size cannot be decided *a priori* but it can be observed at the end of the sampling process.

The steps explained here, except for Poisson and Bernoulli Sampling, are repeated *m* times, where *m* is the number of tuples to be added to the sample. The ratio $\frac{n}{m}$ is the *sampling scale factor* and it is used to estimate real values. As an example, if *n* = 1000 and *m* = 500, we are sampling at 50% and, therefore, the sampling scale factor is 2. So, if a sum operation gives 30 as a result against the sample, this value must be scaled and, then, the exact sum value is approximately $2 \times 30 = 60$.

Variance and standard deviation are used to bound the error of the approximate answer. Indeed, a complete approximate answer has the form: $\alpha \pm \delta$ with probability *p*%, where α is the approximate answer, δ is the confidence interval, and *p* is the confidence degree. For example, 100 $\pm$10 with *p*=90% means that there is the 90% of probability that the exact value of the aggregate query falls in the range [90, 110].

3 Proposed method

Let Q_a be an aggregate query over the numerical attribute *a*.
The coefficient of variation of *a* measures the spread of values around their average and is computed as

$$cv(a) = \frac{standard\ deviation(a)}{average(a)}$$

In literature, a coefficient of variation equals or higher than 1 implies a high variance [5]. On the other hand, *cv*(*a*) = 0 if and only if the standard deviation is 0, that is all the values are equal. If the coefficient of variation is 0, we may sample only one value (*i.e.*, only the first, for example). As the coefficient of variation approaches to 1, we need to sample more and more values to obtain representative data and, then, minimize the approximation error during analytical processing.

We assume the minimum sample size is the 10% of the population, while the maximum sample size is the 90%. The minimum sample size can be adopted when the coefficient is close to 0, while the maximum when the coefficient is equals to 1 or higher.

Formally, we introduce the sampling function *sp* as a mapping of the coefficient of variation to the range [10, 90], which indicates the sampling percentage

$$sp\colon cv(a) \in \mathbb{R} \longrightarrow [10{,}90].$$

The sampling function is computed as follows

$$sp(x) = \begin{cases} 10, & x \leq 0.1 \\ \omega, & 0.1 < x < 1 \\ 90, & x \geq 1 \end{cases}$$

where

$$\omega = \frac{x-0.1}{1-0.1}(90-10) + 10.$$

So, each aggregate query involving the numerical attribute *a* requires a sampling percentage equals to *sp*(*x*), where *x* is the coefficient of variation of *a*. Given the sampling percentage, we can derive the sampling scale factor.

An implementation of this sampling method is *permic* (Perturbation-based Method for Inference Control), that can be found at [7]. This implementation does not perform a data reduction but loads into memory the percentage of data on the basis of the computed coefficient of variation.

3.1 Experimental result

The experiment generates four dataset having mean $\mu = 50$ and variance $\sigma^2 \in \{1, 10, 100, 1000\}$. Since we are not interested in answering times, the cardinality of the datasets is irrelevant in this context. The aggregate query is a *sum* operation over the entire dataset. The result of the experiment is reported in Table 1.

Variance	Approximate	Exact	Relative error (%)
1	50030	49998	0,06
10	49940	49895	0,09
100	49021	49923	1,80
1000	49875	49605	0,54

Table 1. Experimental data.

Experimental data show that the relative error is quite constant and lower than 2% but, most important, it does not depend on the variance. So, we can conclude this preliminary study observing that, using a sample size varying in reference to the variation, we can obtain a representative data synopsis such that approximation error is minimized in analytical processing.

4 Conclusion

In the paper, a sampling method for approximate query processing is proposed. The method is *adaptive*, because the sample is established at run-time on the basis of the statistical profile of the database. In detail, the method relies on the coefficient of variation that determines the sample size in order to best approximate exact answers of aggregate queries in analytical processing. Experimental results show that the relative error does not depend on the variance. This method has been successfully implemented in an approximate query answering system that performs an *on-line* sampling, avoiding the necessity of the data reduction process.

References

1. Madigan, David, and Martha Nason. "Data reduction: sampling." *Handbook of data mining and knowledge discovery*. 2002. 205-208.
2. Cormode, Graham, et al. "Synopses for massive data: Samples, histograms, wavelets, sketches." *Foundations and Trends in Databases* 4.1–3 (2012): 1-294.
3. Denning, Dorothy, and Jan Schlorer. "Inference controls for statistical databases." *IEEE Annals of the History of Computing* 16.07 (1983): 69-82.
4. Liu, Qing. "Approximate Query Processing." (2009): 113-119.
5. Savatovic, Anita, and Mejra Cakic. "Estimating Optimal Checkpoint Intervals Using GPSS Simulation." (2007).
6. Van Belle, Gerald, and Donald C. Martin. "Sample size as a function of coefficient of variation and ratio of means." *The American Statistician* 47.3 (1993): 165-167.
7. https://github.com/francescoditria/permic