IEEE TRANSACTIONS ON

KNOWLEDGE AND DATA ENGINEERING

A publication of the IEEE Computer Society

MARCH / APRIL 1999 VOLUME 11 NUMBER 2 ITKEEH (ISSN 1041-4347)

REGULAR PAPERS

CORRESPONDENCE

Dr. Naphtali Rishe
Florida International University
School of Computer Science
Southwest 8th St. and 107th Avenue
University Park
Miami, FL 33199

IEEE COMPUTER SOCIETY

http://computer.org • tkde@computer.org

A Hybrid Estimator for Selectivity Estimation

Yibei Ling, Wei Sun, *Senior Member, IEEE*,
Naphtali D. Rishe, *Member, IEEE Computer Society*, and Xianjing Xiang

Abstract—Traditional sampling-based estimators infer the actual selectivity of a query based purely on runtime information gathering, excluding the previously collected information, which underutilizes the information available. Table-based and parametric estimators extrapolate the actual selectivity of a query based only on the previously collected information, ignoring on-line information, which results in inaccurate estimation in a frequently updated environment. We propose a novel hybrid estimator that utilizes and optimally combines the on-line and previously collected information. Theoretical analysis demonstrates that the on-line and previously collected information is complementary and that the comprehensive utilization of the on-line and previously collected information is of value for further performance improvement. Our theoretical results are validated by a comprehensive experimental study using a practical database, in the presence of insert, delete, and update operations. The hybrid approach is very promising in the sense that it provides the adaptive mechanism that allows the optimal combination of information obtained from different sources in order to achieve a higher estimation accuracy and reliability.

Index Terms—Hybrid estimator, sampling estimator, parametric estimator, table-based estimator, query optimization, estimation accuracy, estimation reliability.

1 Introduction

As a rather efficient, accurate, reliable means of determining the optimal query execution plan among many equivalent query execution plans of different costs, sampling-based estimators have received extensive attention and well studied [4], [5], [14], [15], [16], [17], [19], [18], [24], [28], [29], [30], [25], [26], [31], [10], [34], [37]. The advantages of sampling-based estimators lie in their good reliability in truthfully reflecting runtime data distributions, and their robustness in the presence of correlated data.

The benefits of sampling-based etimators, however, come at a price. Sampling-based algorithms conduct runtime information gathering, and extrapolate the resulting size from the sampled data. As a result, a certain amount of runtime sampling overhead has to be incurred whenever a size estimation has to be made. This overhead adds to the response time of query processing. In addition, the sampled information is completely volatile: The sampled information obtained for the current query has to be discarded, and can not be reused for subsequent queries. Research on sampling-based methods places an emphasis on the strategy for minimizing the sample size while satisfying the required estimation accuracy at the given confidence level [16], [15], [14], [19], [18], [24], [28], [29], [30], [25], [26], [31], [34], [37]. However, reducing the runtime sampling overhead and increasing the estimation accuracy seem to be inherently contradictory: generally speaking, a higher estimation accuracy requires a larger sampling size, which in turn results in a larger I/O overhead. Therefore, a balanced trade-off between estimation accuracy and runtime sampling overhead must be made.

Parametric estimation methods [1], [2], [7], [8], [12], [13], [22] use certain statistical functions to describe the data so as to provide estimation by evaluating its approximating function when a query is given. Different parametric models including uniform distribution, normal distribution, neural learning networks and regression [25], [36], [22], [6] have been proposed to approximate the actual data distributions. The benefit of parametric estimators lies in the efficient estimation computation.

Table-based estimators [9], [32], [20], [33], [21] use the stored summary statistics to estimate the resulting size of a query, and can provide accurate estimations in a retrieval-intensive or retrieval-only environment, but at the expense of the runtime overhead for maintaining summary statistics.

Both parametric and table-based estimators are widely used by commercial database system such as Oracle, Sybase, SQL Server, Ingres, and DB2. However, parametric and table-based estimators may perform rather poorly in the presence of frequent update operation, since the frequent update will substantially change the underlying database, and make the established table and/or model rapidly outdated. To maintain the estimation accuracy, the stored summary statistics (table-based estimator) or the approximating statistical model (parametric estimator) needs to be computed periodically by using the up-to-date information about the underlying data distribution.

In principal, parametric and table-based estimators infer the resulting size of a query based purely on the previously collected information, ignoring on-line information; while

- *Y. Ling is with the Bell Communications Research Laboratory, Bellcore, 445 South St., Morristown, NJ 07960. E-mail: lingy@bellcore.com.*
- *W. Sun and N.D. Rishe are with the School of Computer Science, Florida International University, Miami, FL 33199. E-mail: {weisun, rishen}@cs.fiu.edu.*
- *X. Xiang is with the Department of Biostatistics, Novartis Pharmaceuticals, Summit, NJ 07901. E-mail: xiaojing.xiang@pharma.novartis.com.*

Manuscript received 16 Aug. 1996; revised 28 May 1998.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number 104340.

sampling-based estimators extrapolate the actual selectivity based purely on the on-line information gathering, excluding the previously collected information. In an effort to overcome these deficiencies, we introduce a hybrid model that fully utilizes the on-line information obtained by a sampling-based estimator and the previously collected information obtained by a parametric/table-based one, partaking of the merits of the participating estimators. Our approach, in its style, is similar to that proposed independently by Haas and Swami [17] for reducing the estimation variability by using previously stored AFV statistics and on-line information from sampling.

The comprehensive utilization of the on-line and previously collected information makes the hybrid model outstanding in the following respects, which are substantiated by our comprehensive experimental study as reported in this paper:

- **Higher Reliability**: Our analysis shows that the mean-squared error of the hybrid estimator is strictly smaller than those of the two participating models being utilized in the hybrid model. As a result, the hybrid model is of a better coverage [27] than the two participating estimators.
- **Higher Efficiency**: We show that, as compared with pure sampling-based estimators, the hybrid estimator can make a substantial saving in sample size when the same estimation accuracy is required.
- **Higher Accuracy**: The hybrid estimator provides a more accurate estimation. As compared with pure sampling-based estimators (no matter which sampling method is used), the estimation accuracy can be significantly improved by the hybrid estimator when the same sample size is used.

The paper is organized as follows: Section 2 outlines the hybrid estimator, and provides theoretical results regarding the superiority of the hybrid estimator over its two participating estimators as well as a quantitative analysis of the additional benefit gained by the hybrid estimator in comparison with a pure sampling-based estimator. Section 3 presents an experimental study of the hybrid estimator with different selectivities and sample sizes. Section 4 presents the performance study of the hybrid estimator based on the practical movie database in the presence of update operations. Section 5 concludes this paper.

2 Hybrid Estimator

Let's start with the sampling-based approach. The main idea behind the sampling techniques is to repeatedly select a tuple randomly from a table against the given query predicate, then make inference about the actual selectivity by using the estimated selectivity obtained from the sampled data. Its procedure can be formalized as follows.

Let $f()$ be the *characteristic function* of a selection predicate, and let x_i be a tuple. Given a selection, $f()$ can be defined as follows:

$$y_i = f(x_i) = \begin{cases} 1 & \text{if } x_i \text{ satisfies the selection predicate} \\ 0 & \text{otherwise} \end{cases}$$

$$\widehat{p_n} = \sum_{i=1}^{n} y_{r_i} \; / \; n = \sum_{i=1}^{n} f(x_{r_i}) \; / \; n$$

where an index r_i is a random integer between 1 and k, k is the total number of tuples, and n is the sample size. Based on the sampled information, we can infer that $\widehat{p_n}$ in the above formula is an approximation of the actual selectivity p. Thus, the actual number of tuples satisfying the selection can be approximated as $k \times \widehat{p_n}$.

We now give the details of the hybrid estimator and its theoretical analysis as follows:

$$\breve{p}_n = t \cdot \widehat{p_n} + (1 - t) \cdot \tilde{p} \tag{1}$$

where $\widehat{p_n}$ is the estimated selectivity from a pure sampling-based method, the subscript n denotes the sample size, $\tilde{p}$ is the estimated selectivity obtained by a parametric estimator or by a table-based estimator, and the parameter t is in the range [0, 1]. It is clear that the estimated selectivity of the hybrid estimator $\breve{p}_n$ is obtained by a linear combination of information obtained from the two participating estimators, which makes the hybrid estimator distinguished from the traditional sampling-based, parametric and table-based estimators.

To assess an estimator, we use the mean-squared error (*mse*) to quantify the performance of an estimator as follows:

$$mse = E(\bar{p} - p)^2 = \sum_{i=1}^{n} (\bar{p}_i - p)^2 \; / \; n$$

where $\bar{p}_i$ is an individual estimated selectivity by an estimator, p is the actual total selectivity with respect to the given query, and n is the sample size.

The *mse* of an estimator represents its estimation accuracy as well as reliability; the smaller the *mse* is, the better the estimator is. It is known that the *mse* for a sampling method is $p \cdot (1 - p)/n$ [11]. The *mse* for an estimator using a table-based or a parametric method is $(\tilde{p} - p)^2$, and $\tilde{p}$ is unchanged with respect to a given query, until the parametric or table-based estimator has been recomputed using the up-to-date information.

Different t values represent different weights being imposed on the two participating estimators. In the extreme cases that $t = 1$ or $t = 0$, the hybrid model is reduced to a pure sampling-based estimator or a parametric/table-based estimator, respectively. In other words, a sampling-based or parametric/table-based estimator is only a special case of the hybrid estimator. The following Theorem illustrates the existence of an optimal value for t and shows how to determine the value for t that optimally combines the participating estimators, resulting in a performance improvement over its two participating estimators.

THEOREM 1. *The optimal value for t, denoted as* t_n^*, *is*

$$\frac{(\tilde{p}-p)^2}{p\cdot(1-p)/n+(\tilde{p}-p)^2}.$$

And the mse for the hybrid estimator with the optimal parameter t_n^* *is smaller than either of the two participating estimators when* $0<p<1$, *and* $p\neq\tilde{p}$, *namely,*

$$E\left(\breve{p}_n^*-p\right)^2 < min\left\{\frac{p\cdot(1-p)}{n},\ (\tilde{p}-p)^2\right\}$$

where $\breve{p}_n^*$ *is the* $\breve{p}_n$ *with* $t=t_n^*$.

The proof of Theorem 1 and its detailed derivation can be found in Appendix A. In Fig. 1, the relationship between the *mse* of the hybrid estimator and the parameter t is presented. To illustrate how Theorem 1 works, let's assume that the total selectivity is $p = 0.2$, the sample size n is 50, and $\tilde{p} = 0.25$. The relative estimation error by using the previous information is assumed to be $|p-\tilde{p}|/p = 25$ percent. Then, according to Theorem 1, $t_n^* = 0.438596$.

$$E(\breve{p}_n^*-p)^2 = 0.00140351,$$

which is smaller than the *mse* of the sample method $p\cdot(1-p)/50 = 0.0032$, and the *mse* of the estimator using the previous information $(\tilde{p}-p)^2 = 0.0025$.

It is interesting to observe from Theorem 1 that

- When sample size n is large, $p\cdot(1-p)/n \ll (\tilde{p}-p)^2$, so the optimal $t_n^*\to 1$, as $n\to\infty$. In this case, the information obtained from a pure sampling-based is heavily weighted in the hybrid estimator, the hybrid estimator is largely contingent on the on-line information gathering;
- When the sample size n is small and the previously collected information is accurate ($|p-\tilde{p}|^2$ is small), $(\tilde{p}-p)^2 \ll p\cdot(1-p)/n$, so the optimal $t_n^*\to 0$, as $n\to 0$. In this case, the previously collected information predominates the on-line information, constituting the principle aspect of the hybrid estimator.

The value for t_n^* indicates that the hybrid estimator takes $100\cdot(t_n^*)$ percent of the on-line sampling information, and $100\cdot(1-t_n^*)$ percent of the previously collected information, forming an optimal mixture of the information from the different sources, leading to a higher estimator accuracy and reliability.

Determining the optimal t_n^* in Theorem 1 requires knowledge of the actual selectivity p, which in practice is unknown. Therefore p needs to be estimated. In Section 3, we calculate t_n^* by replacing p by $(\widehat{p_n}+\tilde{p})/2$, and show that this substitution could result in a substantial performance improvement over a pure sampling-based estimator. It should be emphasized that the technique we used here for substituting the actual p in t_n^* is similar to those in sampling-based algorithms [16], [15], [17], [18], [28], [29], [30], [26] for replacing the actual selectivity p with the estimated $\widehat{p_n}$ in determining the termination condition of sampling. At the end of Section 4, we give guidelines for the substitution that ensures the superiority of the hybrid estimator over its two participating estimators.

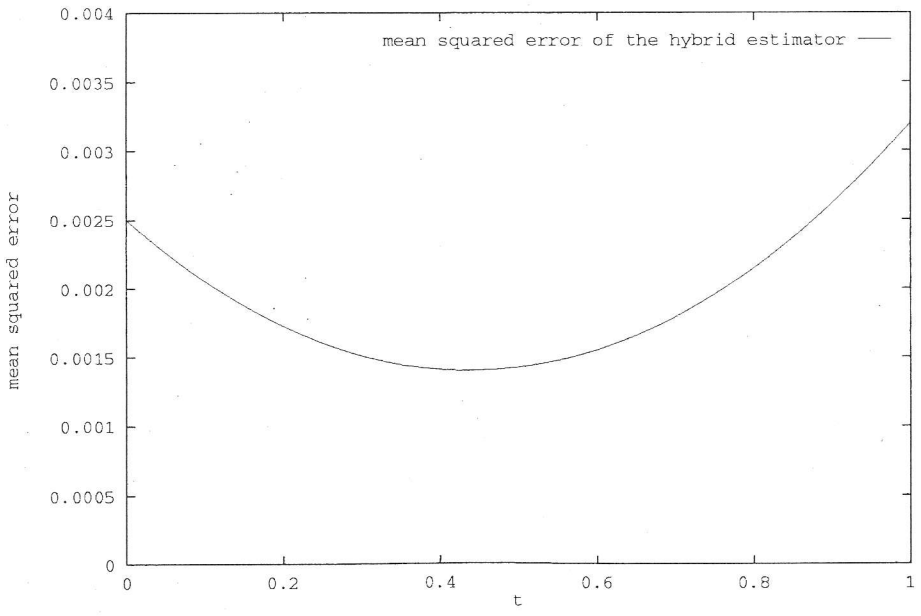

Fig. 1. The *mse* of the hybrid estimator vs. parameter *t*.

Theorem 2 guarantees that the hybrid estimator provides a more accurate and reliable size estimation in any mathematical measurements than its two participating estimators.

THEOREM 2. *For any*

$$\alpha \geq 1, \text{ if } E|\breve{p}_n^* - p|^\alpha = min_{0\leq t\leq 1} E|\breve{p} - p|^\alpha,$$

then

$$E|\breve{p}_n^* - p|^\alpha \leq min(|\tilde{p} - p|^\alpha, E(\hat{p}_n - p)^\alpha).$$

The proof of Theorem 2 can be found in Appendix A. Theorem 2 is very strong in the sense that it implies that the estimation error of the hybrid estimator could be smaller than that of the two participating estimators in any meaningful mathematical measurements.

Theorem 3 provides a quantitative account of the benefit of the additional information to the hybrid estimator, in comparison with a pure sampling-based estimator.

THEOREM 3. *Let m be the sample size for a pure sampling-based estimator, and n the sample size for the hybrid estimator. Assume that both estimators have equal mse, that is*

$$E(\hat{p}_m - p)^2 = E(\breve{p}_n^* - p)^2. \quad (2)$$

Then, we have $m - n = m \cdot \alpha$, *where*

$$\alpha = \frac{p \cdot (1-p)/n}{p \cdot (1-p)/n + (\tilde{p} - p)^2},$$

and $(m - n)$ *represents the sample size saved by the hybrid method over a pure sampling estimator.*

Proof of Theorem 3 can be found in Appendix A. Note that α is actually a ratio of the sample size saved by the hybrid estimator to the sample size required by a pure sampling-based estimator. Theorem 3 illustrates that accurate previous information ($|\tilde{p} - p|$ is small, and α is close to 1) could result in a substantial sampling reduction. The following example is given to illustrate how to quantify the benefit of the hybrid estimator. Let's assume that $\alpha = 0.9$, based on Theorem 3, the benefit, which is expressed in terms of the sampling reduction, of the hybrid estimator can be written as

$$m - n = m \cdot 0.9, \quad then \quad m = 10 \cdot n$$

that is, to achieve the same estimation accuracy, the sample size required by a sampling-based estimator is 10 times as much as that used for the hybrid estimator. The above example reveals that the availability of the previous knowledge, even partially outdated due to the presence of updates, is valuable, and can be used to reduce the sample size and increase estimation accuracy.

3 A Performance Study

In principle, any table-based estimator or parametric estimator can be used as one of two participating estimators for the hybrid estimator to gather the previously stored information. In this performance study, we choose a *self-organizing* model [22], which can be considered as a parametric estimator. The underlying reason for choosing *self-organizing model* is based on our extensive experience with this model.

The *self-organizing* model is based on a neural network learning model which has gained popularity in recent years, and has had immense success in various areas [3], [35], [38], [23]. A neural network model basically represents an approximating function for a cumulative data distribution defined as follows:

$$F(d) = \sum_{x\leq d} f(x)$$

where $f(x)$ is the data distribution representing the number of tuples having value x under the concerned attribute. By the above definition, it is clear that:

1) $F(x)$ is a nondecreasing function defined on the domain, that is, $\forall x_1, x_2 \in dom(x)$, if $x_1 \leq x_2$, then $F(x_1) \leq F(x_2)$, and
2) $F(x)$ is bounded by the actual number of tuples, namely $F(x) \leq |R|$.

Properties 1-2 are invariant regardless of the data distribution under R.X. Given two domain values x_1 and x_2, and $x_1 \leq x_2$, it can be directly observed that $(F(x_2) - F(x_1))$ represents the number of tuples satisfying the selection predicate $(x_1 < R.X \leq x_2)$. A cumulative data distribution under an attribute can be easily constructed by a sequential scan of the whole relation R.

The *self-organizing* model $G(x)$ [22] has been proposed to identify the cumulative data distribution $F(x)$. The self-organizing model is initially established by using the *back propagation learning rule* and training data $F(x)$. The left-hand side of Fig. 2 represents the actual cumulative distribution $F(x)$ derived from a uniform distribution, while the right-hand side represents the approximating function $G(x)$ obtained through the neural network learning process. Fig. 2 shows that the self-organizing model, once trained, can well approach $F(x)$.

Instead of using the cumulative data distribution $F(x)$, we use the approximating function $G(x)$ obtained through the training process to estimate the resulting size of a query. When the self-organizing model is established, the resulting size of the selection query $(x_1 < R.X \leq x_2)$ can be approximated as $G(x_2) - G(x_1)$. The self-organizing model requires neither extra on-line overhead for information gathering, nor the huge space overhead for storing the statistics about the underlying data distribution, and can provide efficient and accurate size estimation.

$$G(x) = 10000 \cdot \sin(0.920305 \cdot z + 0.979493 \cdot z^2 - 2.749408 \cdot z^3 + 2.419611 \cdot z^4),$$

where $z = x/1000$. Like neural network training, to obtain $G(x)$ initially requires moderate non-runtime overhead in identifying the optimal coefficients, then the self-organizing model can adjust its weights based on the query feedback information [6], [22]. The distinction between the self-organizing model and neural network model is that the self-organizing model can be automatically adaptive to the constantly changing data distribution using a query feedback mechanism with negligible computation overhead. As

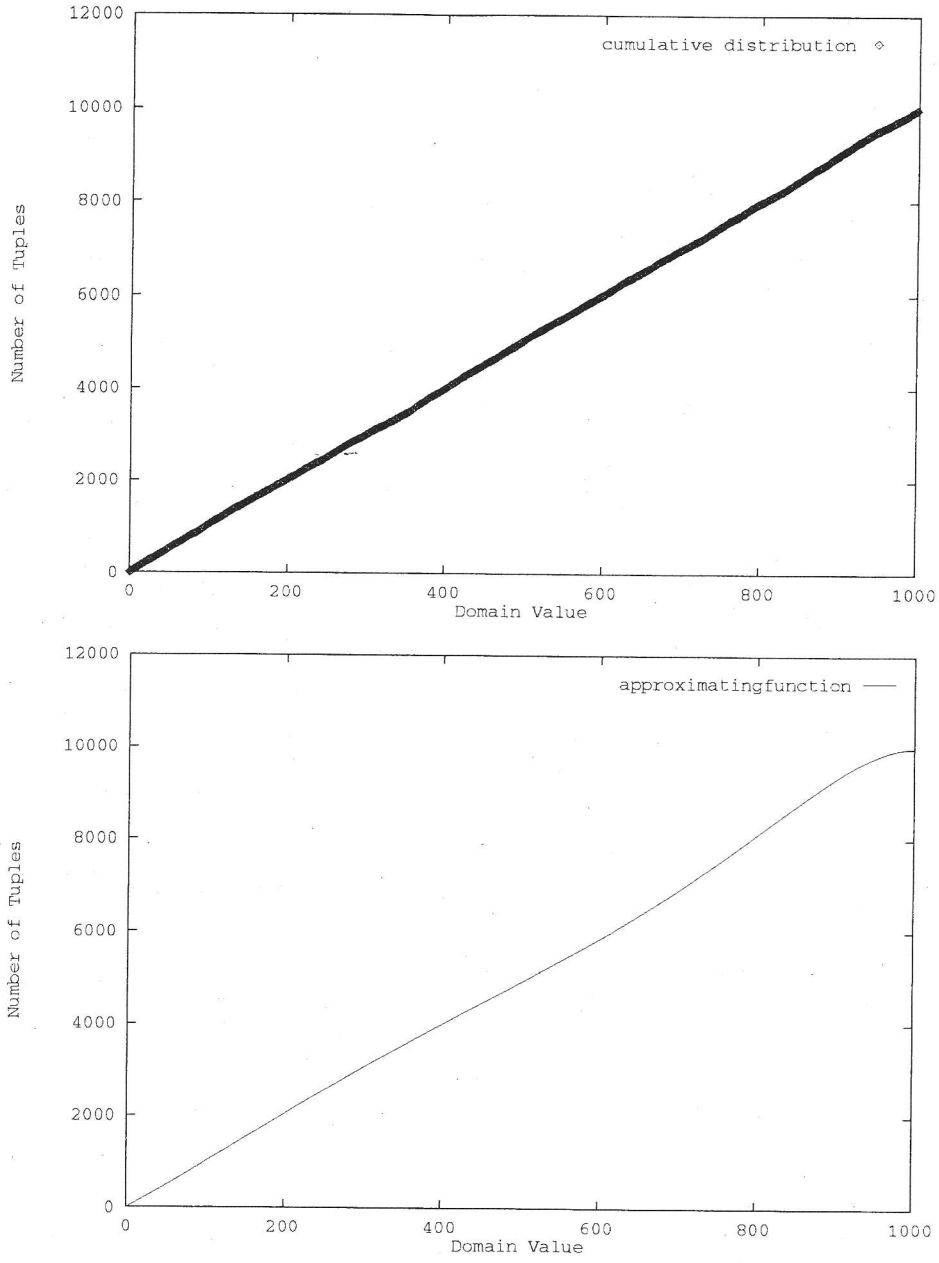

Fig. 2. Cumulative data $F(x)$ and approximating function $G(x)$.

a result, the performance of the self-organizing model is rather persistent in the presence of update. In the following experimental study, we use the self-organizing model and random sampling estimator as the two participating estimators of the hybrid estimator. It is worth noting that the hybrid estimator does not distinguish which sampling method and parametric/table-based method are being used.

To fairly evaluate the performance of an estimator, we use the absolute relative estimation error (*aree*, for short) and standard estimation deviation (*sed*, for short) defined as follows:

DEFINITION 1:

$$\text{aree}(\overline{p}, p) = \sum_{i=1}^{n} \left|\overline{p}_i - p\right| / (n \cdot p)$$

$$\text{sed}(\overline{p}, p) = \sum_{i=1}^{n} \left(\overline{p}_i - p\right)^2 / (n \cdot p^2)$$

Notice that the *sed* of an estimator is the corresponding *mse* value amplified by a factor $1/p^2$. It is known that p is the actual selectivity with respect to a given query predicate, and is a constant in our analysis. Therefore, theoretical conclusions obtained are applicable with impunity.

The $aree(\overline{p}, p)$ denotes the average relative estimation error obtained by using the estimator $\overline{p}$ over the population with the total selectivity p. For example, if $\overline{p} = \tilde{p}$, then the $aree(\tilde{p}, p)$ represents the average relative estimation obtained using the self-organizing estimator, $\overline{p_i}$ represents an individual estimated selectivity obtained by using an estimator.

To investigate the performance of the hybrid estimator at the different selectivities, we design the experimental procedure described as follows:

- the underlying data is generated from a uniform distribution ranging from 0 to 1,000;
- the self-organizing model, as plotted in the right-hand side of Fig. 2, is constructed;
- the queries which produce a fixed selectivities (0.8 percent, 5 percent, and 10 percent) are generated (for uniform data distribution, the total selectivity p can be controlled by choosing the range of query);
- the estimated selectivity $\widehat{p_n}$ obtained from random sampling with the different sample sizes are collected;
- the estimated selectivity $\tilde{p}$ obtained from the self-organizing model is collected, the weights of the self-organizing model are adjusted based on the query feedback result;
- the optimal value t_n^* is calculated based on Theorem 1 by replacing p with $(\widehat{p_n} + \tilde{p})/2$;
- the estimated selectivity of the hybrid estimator $(\breve{p}_n^*)$ is calculated;
- the above procedure is repeated, the *aree* and *sed* of the hybrid estimator and random sampling are collected and plotted in Fig. 3 and Fig. 5, in which the relationship between the estimation error of the estimators and the sample size taken is illustrated.

Notice that every point in Fig. 3 and Fig. 5 represents the average measurement obtained from 100 repetitions under the different selectivity p, the left-hand and right-hand sides represent the *aree* and *sed* of the estimators. The captions of Fig. 3 and Fig. 5 contain the corresponding $aree(\tilde{p}, p)$ of the self-organizing model since the the self-organizing model bears no relation to the sample size taken.

Comparing the performance of the hybrid estimator with that of random sampling yields an interesting pattern: The hybrid estimator has gained a substantial performance improvement over random sampling when the sample size is small; this improvement becomes gradually insignificant when the sample size becomes large. This observation agrees with theoretical findings revealed in Theorem 1.

Fig. 6 demonstrates the performance difference between the hybrid estimator and its two participating estimators over the different selectivities when a fixed sample size (200) is used. The hybrid estimator generally gives a better result than either the sampling-based and self-organizing estimators over the range of given selectivity. Observed that the performance of the self-organizing model is relatively insensitive to the actual selectivity. The self-organizing model is more accurate and reliable than the hybrid estimator when the selectivity is small (less than 8 percent), departing slightly from Theorem 1. The abnormal situation implies that the substitution of p with $(\widehat{p_n} + \tilde{p})/2$ in t_n^* in Theorem 1 does not always yield the optimal results in the practical application. We will provide the practical guidance of ensuring theoretical superiority of the hybrid estimator over its two participating estimators.

4 A Performance Study Using Practical Data

In this section, we will study the performance of the hybrid estimator using the actual movie database in the presence of update, delete, and insert operations. By courtesy of Professor Gio Wiederhold at Stanford University, we use a movie database that consists of 10,125 movies produced during the years 1900 to 1995. Fig. 7 depicts the cumulative movie data distribution as well as its approximating curve made by the self-organizing model over the entire range from 1900 to 1995.

We generate the random query (a random query means that the range of the query is randomly selected) to collect the estimation measurements of the hybrid estimator and random sampling. Random queries will produce different selectivities, reflecting the working performance in a practical environment.

To quantify the working performance of an estimator, uniform average relative estimation error (*uaree*) and uniform standard deviation (*used*), which represent the average measurement taken over the different selectivities, are used as follows:

Definition 2:

$$\text{uaree}(\overline{p}) = \left(\sum_{i=1}^{n} |\overline{p}_i - p_i| \,/\, p_i \right) / n$$

$$\text{used}(\overline{p}) = \left(\sum_{i=1}^{n} (\overline{p}_i - p_i)^2 \,/\, p_i^2 \right) / n$$

where p_i is the actual selectivity determined by a random query and $\overline{p_i}$ is the estimated selectivity of the actual selectivity p_i using an estimator.

The following experimental study has been designed to show, based on an actual movie database and in the presence of different percentages of update (10 percent, 20 percent, 30 percent, and 50 percent), delete (5 percent, 10 percent, 20 percent, and 40 percent), and insert (5 percent, 10 percent, 20 percent, and 50 percent) operations, the actual *uaree* and *used* for the hybrid estimator and random sampling. The actual *uaree* and *used* of the self-organizing model are also collected and showed in the caption of the corresponding figures.

The experiment study can be described as follows: Delete operations are done by randomly deleting the different percentages of tuples from the existing table, and update and insert operations are done by replacing and inserting

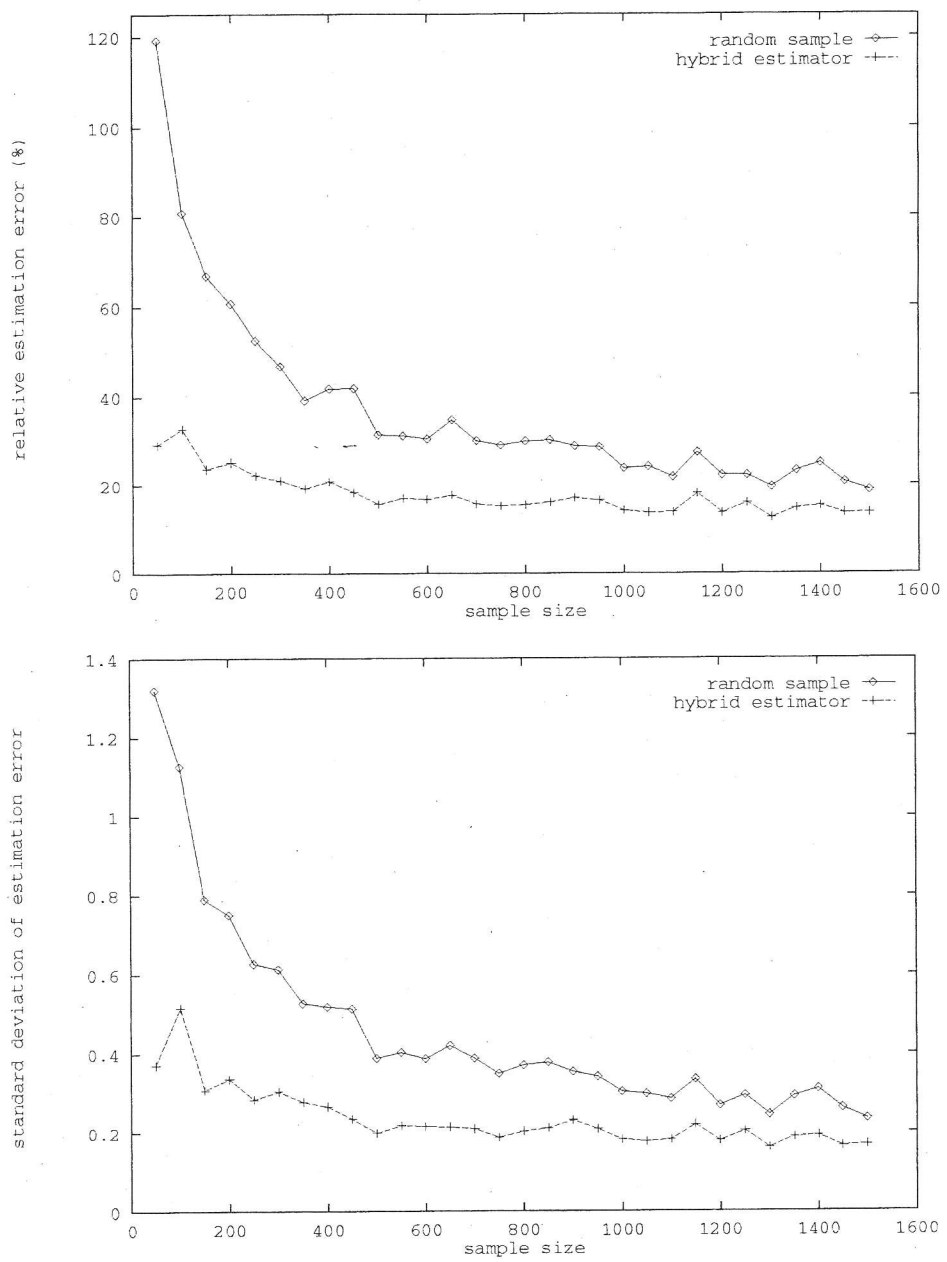

Fig. 3. Estimation accuracy vs. sample size: selectivity p = 0.8 percent (*aree*($\tilde{p}$, 0.008) = 16.08 percent, *sed*($\tilde{p}$, 0.008) = 0.63 percent).

data generated from a uniform distribution into the existing movie database.

It can be observed in Fig. 3 and Fig. 20 that the hybrid estimator performs better than a pure sampling-based estimator in the presence of different percentages of update, delete, and insert operations, that is, the *uaree* and *used* values of the hybrid estimator are generally smaller than those of random sampling, indicating that the experimental study agrees well with theoretical finding.

One point to be noted is that Theorem 1 states that the minimal *mse* of the hybrid estimator can be achieved by choosing:

$$t_n^* = \frac{(\tilde{p} - p)^2}{(\tilde{p} - p)^2 + p \cdot (1 - p) / n},$$

that is, the *mse* of the hybrid estimator is strictly smaller than that of the two participating estimators as long as $0 < p < 1$ and $p \neq \tilde{p}$, representing the optimal combination of the on-line and previously collected information used by the hybrid estimator. However, the total selectivity p is in practice unknown a priori; as an alternative, we replace p with

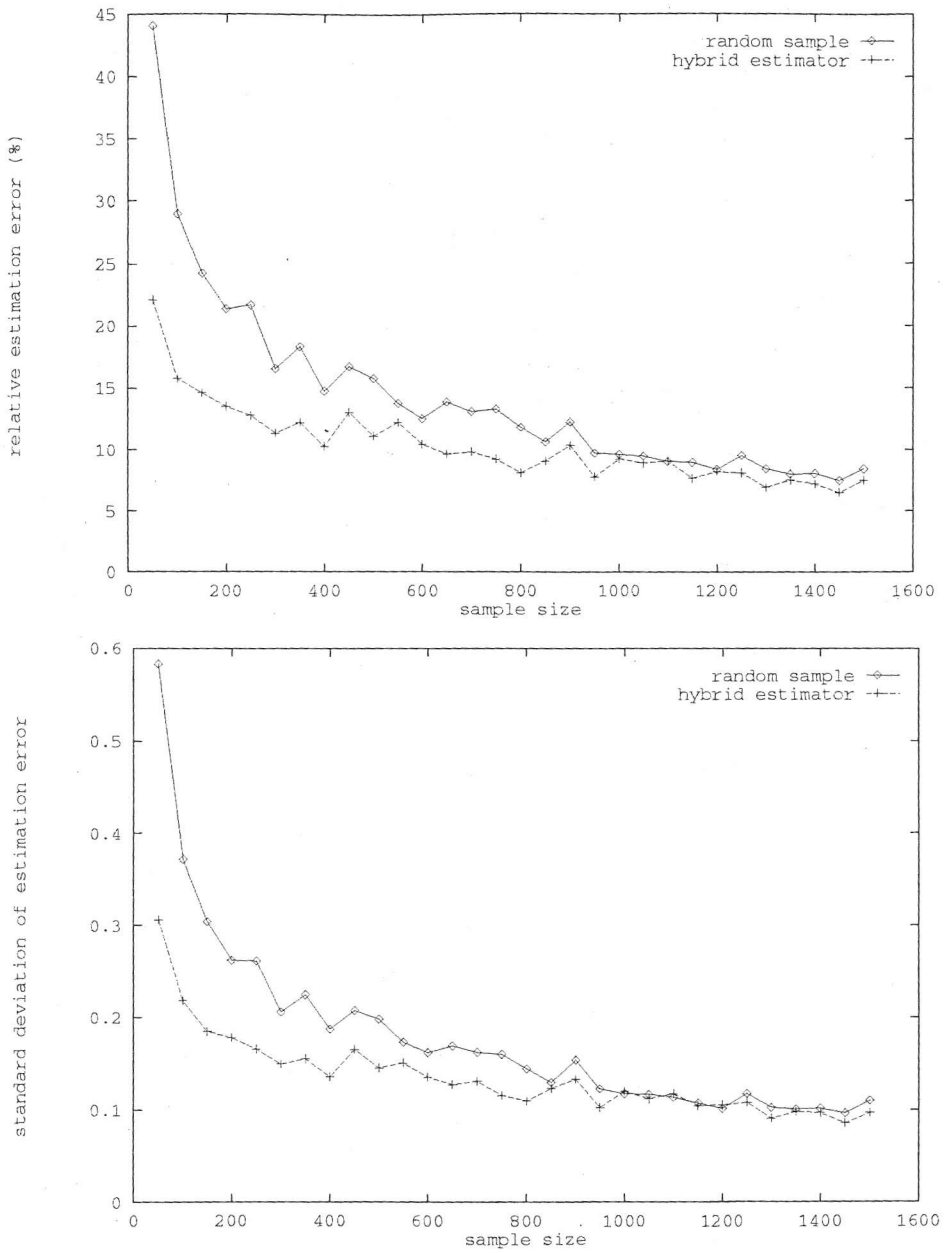

Fig. 4. Estimation accuracy vs. sample size: selectivity p = 5 percent (*aree*($\tilde{p}$, 0.05) = 15.4 percent, *sed*($\tilde{p}$, 0.05) = 0.58 percent).

$$\frac{\widehat{p_n} + \tilde{p}}{2}$$

in the experimental study. Observed from Fig. 3 and Fig. 20 that this substitution sometimes may compromise theoretical superiority of the hybrid estimator to some degree. For instance, in Fig. 6, the *sed* of the hybrid estimator is larger than that of the self-organizing estimator when the sample size is small. Here are guidelines that can be used in practice to obtain t_n^*. We observe that:

1) When the sample size n is large, since $p \cdot (1-p)/n$ becomes small, the selectivity $\widehat{p_n}$ obtained from on-line sampling should be heavily weighted in the substitution of the total selectivity p in the t_n^* formula.
2) When the table-based estimator or parametric estimator has been just updated, or the sample size is small, the selectivity $\tilde{p}$ should be heavily weighted in the substitution of the total selectivity p.

To formulate this idea, the total selectivity p in t_n^* can be expressed as:

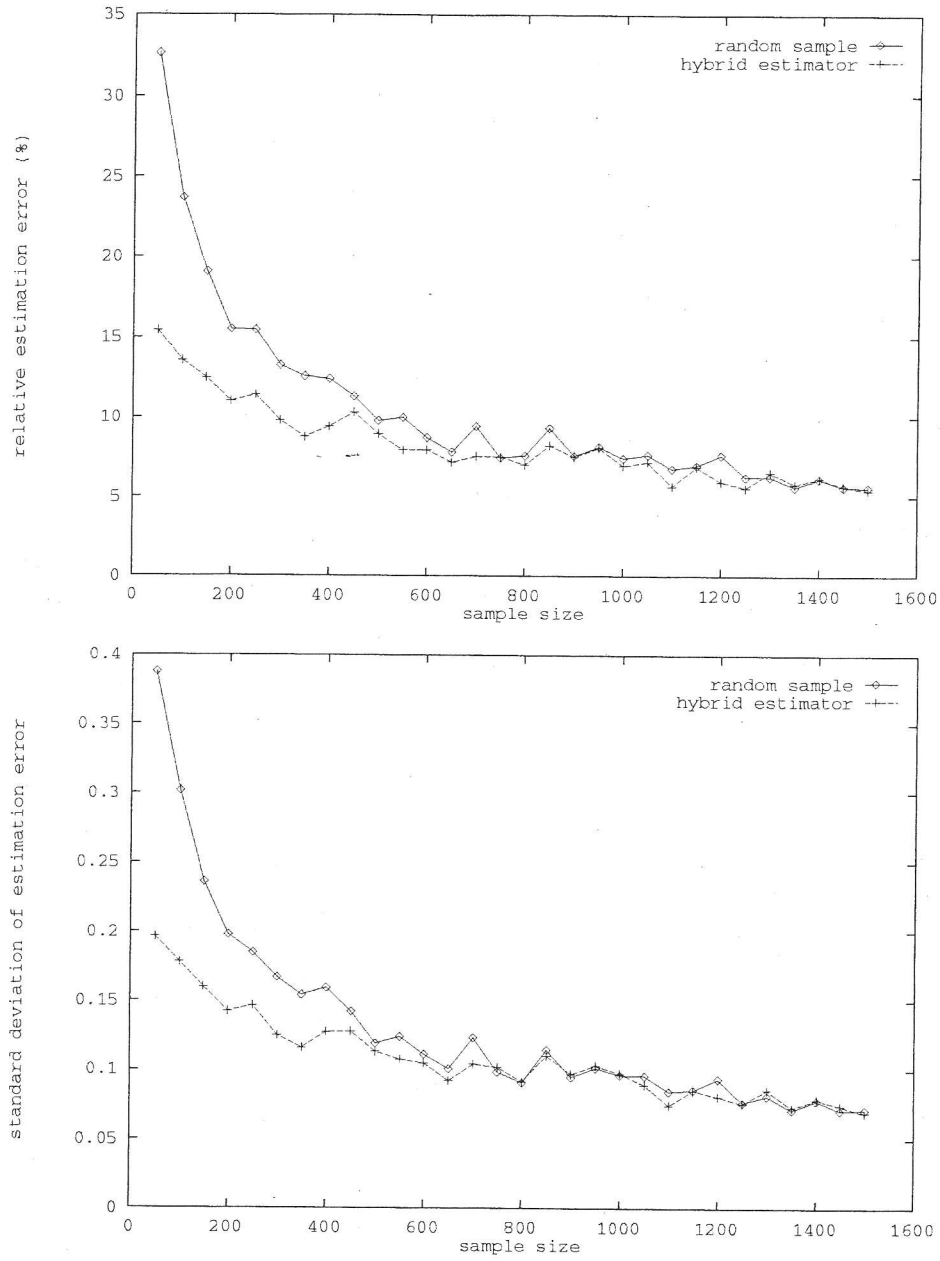

Fig. 5. Estimation accuracy vs. sample size: selectivity p = 10 percent ($aree(\tilde{p}, 0.1)$ = 15.39 percent, $sed(\tilde{p}, 0.1)$ = 0.61 percent).

$$p = (1 - k(n, t)) \cdot \widehat{p_n} + k(n, t) \cdot \tilde{p}, \quad (3)$$

where $k(n, t)$ is a function of the sample size n and the elapsed time t from the last update of the model (parametric or table-based). The function $k(n, t)$ is bounded by 1, and is inversely proportional to the sample size n and the elapsed time t from the last update. The form of the function $k(n, t)$ which could ensure theoretical superiority of the hybrid estimator over its participating estimators requires further investigation and deserves detailed attention.

5 Conclusion

In this paper, we have provided a hybrid estimator that uses the on-line sampling information as well as the previous information furnished by a table-based/parametric method. The contribution of this paper is to show how to utilize information from the different sources and how to determine the optimal combination of the on-line and previously collected information in order to gain the performance improvement.

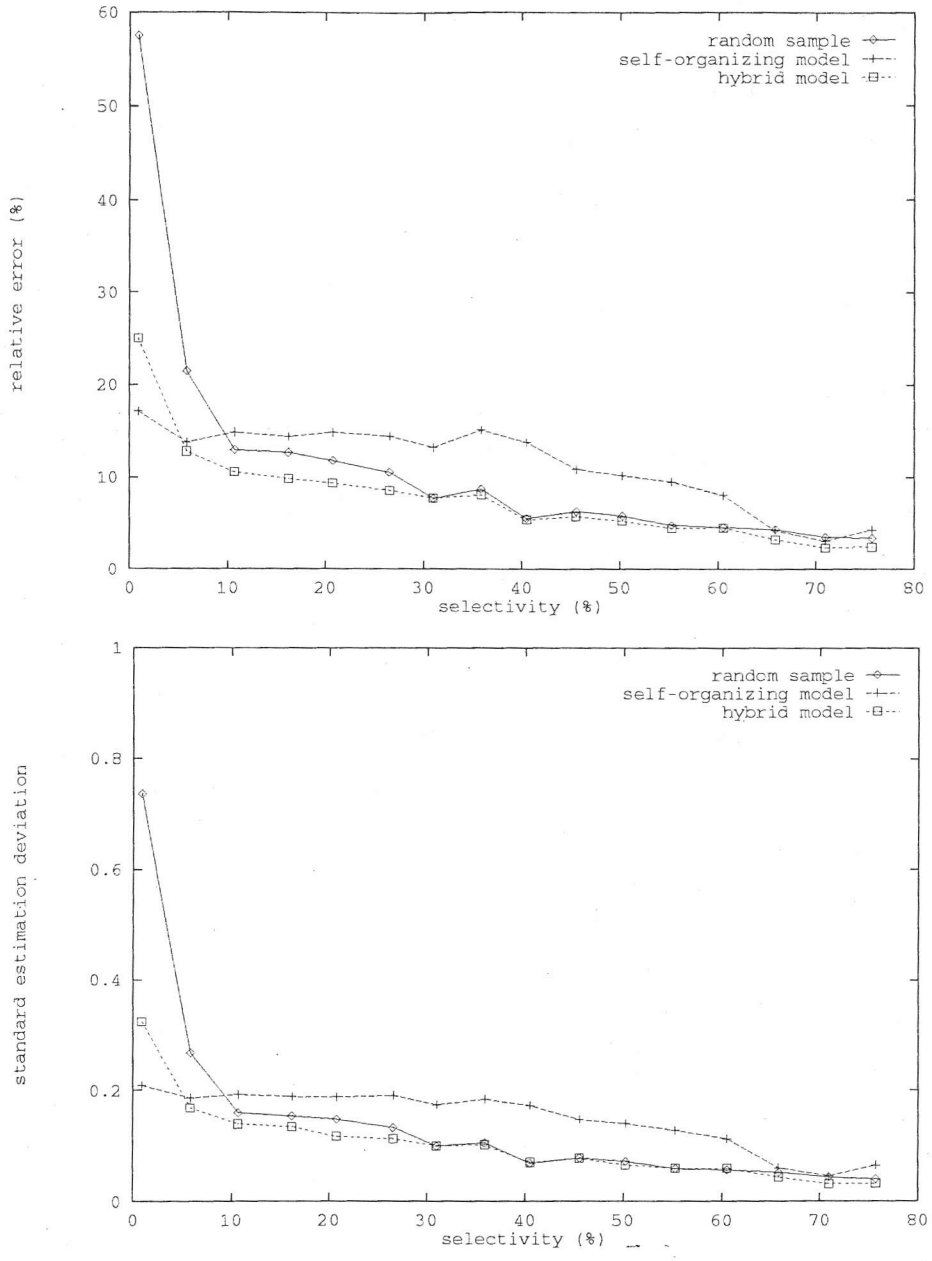

Fig. 6. Estimation accuracy vs. selectivity (percent): (sample size = 200).

Theoretical justification of the superiority of the hybrid estimator, as well as a quantitative analysis of the benefit of the hybrid estimator in terms of estimation accuracy and reliability is presented. The results obtained from the comprehensive experimental study are consistent with theoretical findings.

The hybrid estimator is built based on sampling-based and a parametric/table-based estimator but is independent of the specified estimators used. We have proved that the availability of additional information makes the hybrid estimator theoretically superior to its two participating estimators.

The ability to utilize both on-line and previously collected information makes the hybrid estimator very attractive in a practical setting. Many types of estimators, differing widely in their respective approaches, could be in existence in a database system. The hybrid approach provides an adaptive mechanism that allows the optimal combination of information from heterogeneous source to further increase estimation accuracy and reliability.

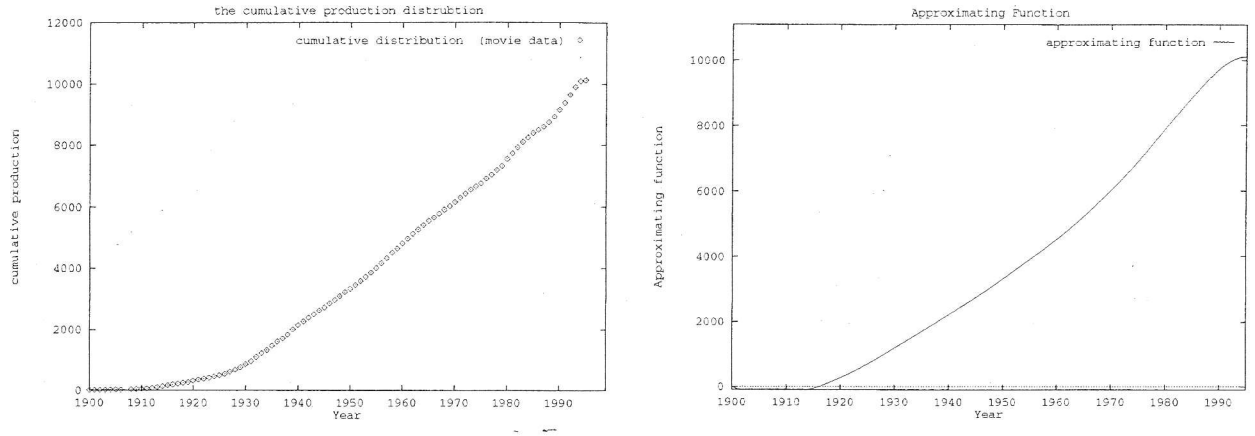

Fig. 7. Cumulative movie data (1900-1995) and its approximating function.

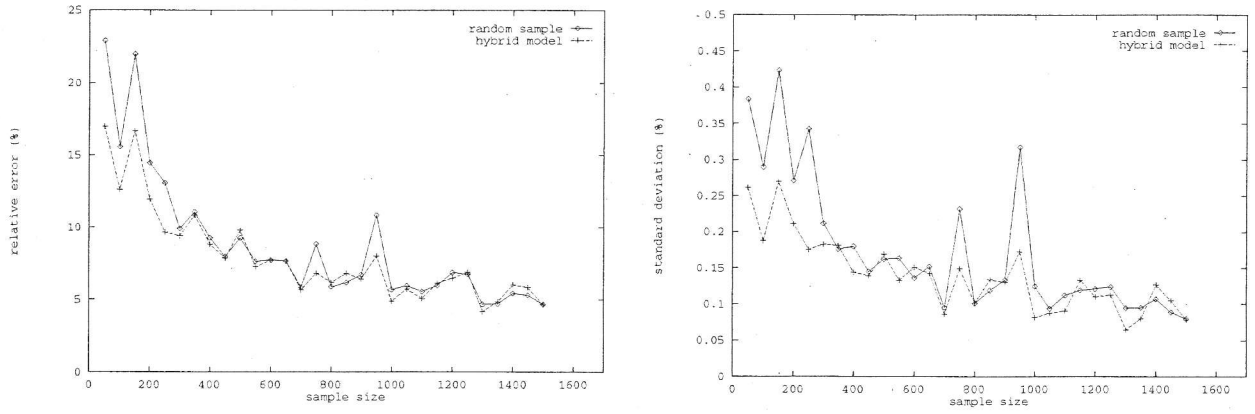

Fig. 8. Performance under the movie data: ($uaree(\tilde{p})$ = 13.39 percent, $used(\tilde{p})$ = 0.14 percent.

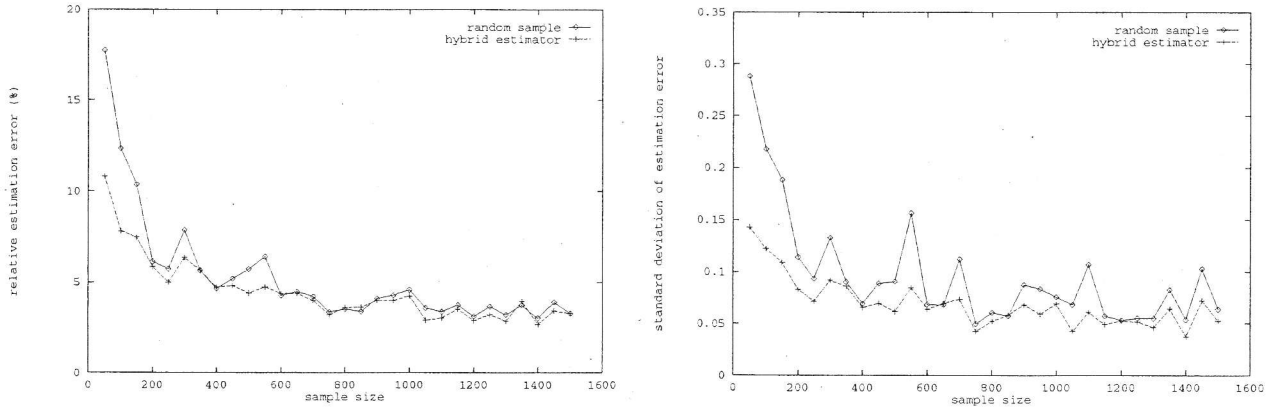

Fig. 9. Performance under deletion (5 percent): ($uaree(\tilde{p})$ = 14.81 percent, $used(\tilde{p})$ = 0.17 percent).

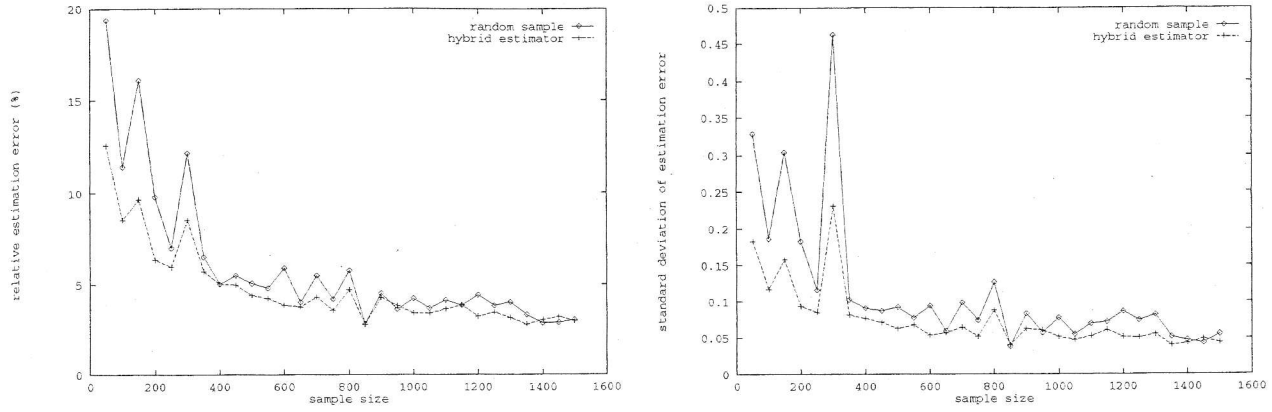

Fig. 10. Performance under deletion (10 percent): ($uaree(\tilde{p})$ = 14.61 percent, $used(\tilde{p})$ = 0.18 percent).

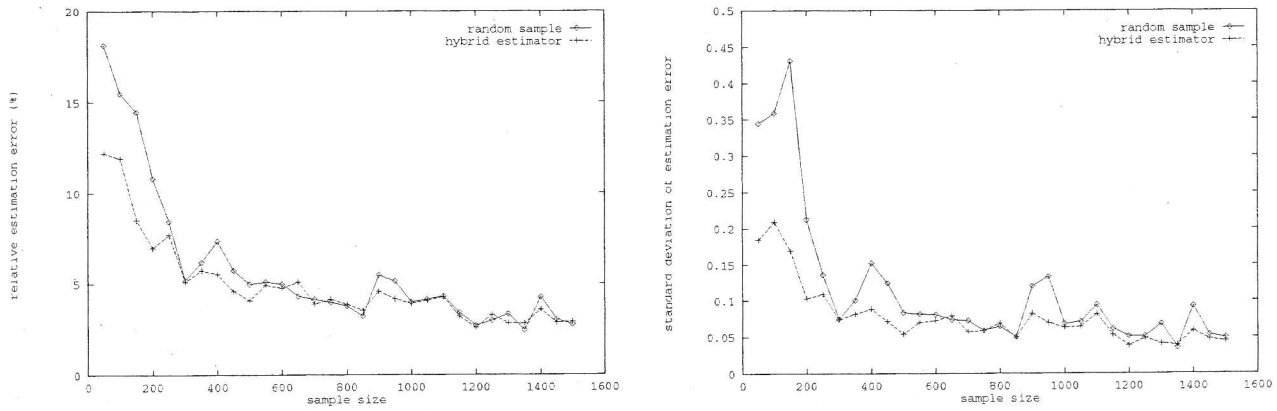

Fig. 11. Performance under deletion (20 percent): ($uaree(\tilde{p})$ = 14.89 percent, $used(\tilde{p})$ = 0.19 percent).

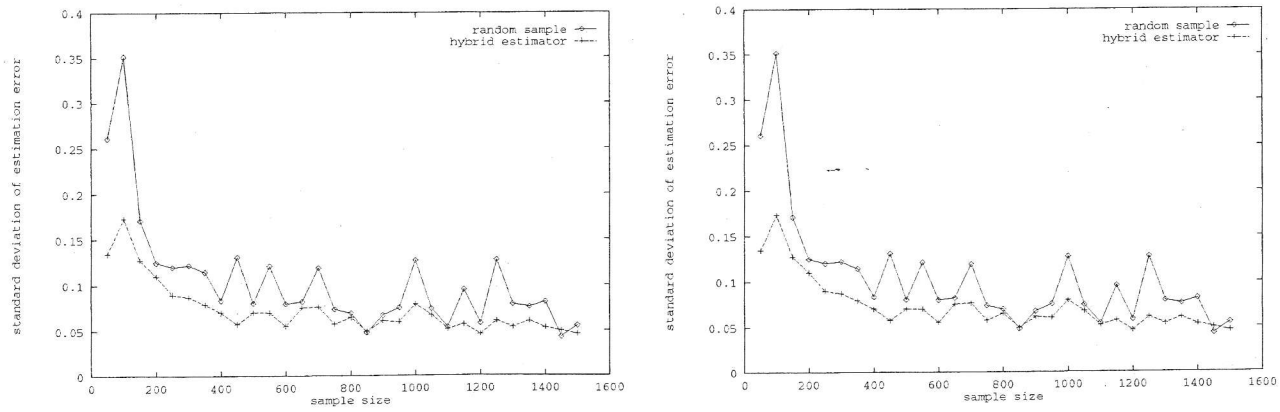

Fig. 12. Performance under deletion (40 percent): ($uaree(\tilde{p})$ = 14.95 percent, $used(\tilde{p})$ = 0.17 percent).

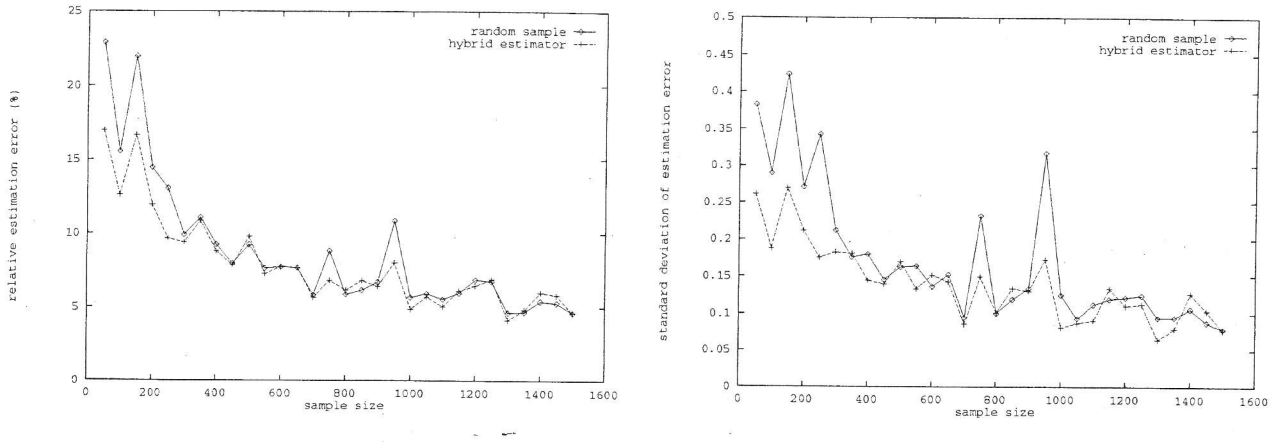

Fig. 13. Performance under update (10 percent): ($uaree(\tilde{p})$ = 21.08 percent, $used(\tilde{p})$ = 0.28 percent).

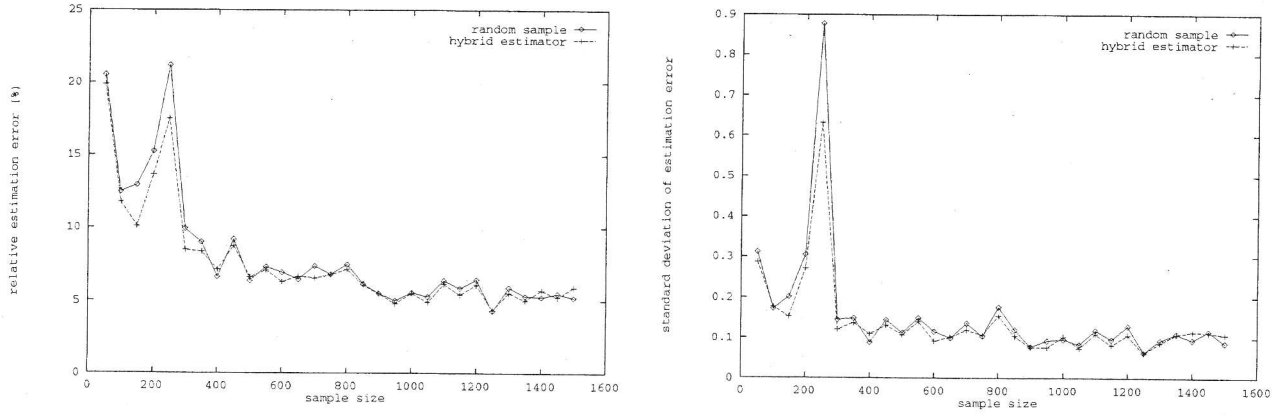

Fig. 14. Performance under update (20 percent): ($uaree(\tilde{p})$ = 22.18 percent, $used(\tilde{p})$ = 0.23 percent).

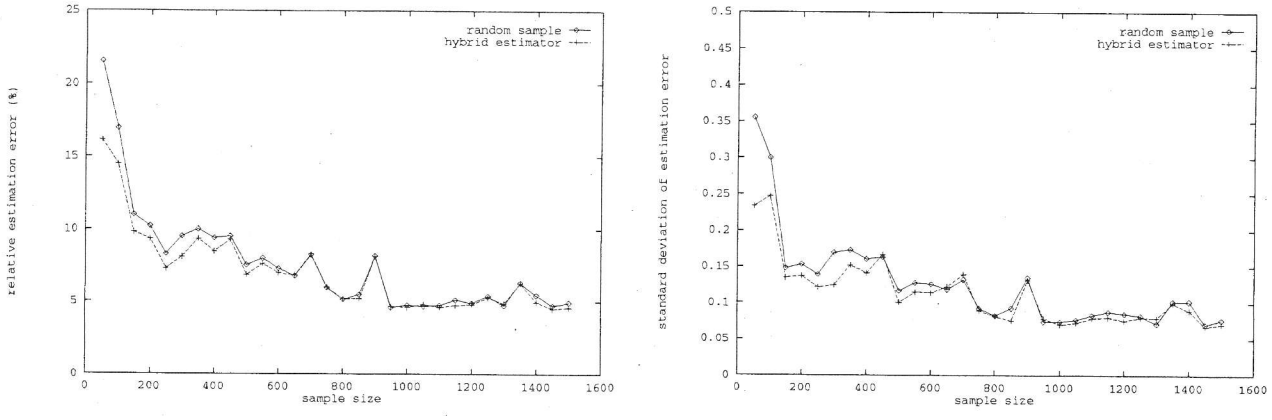

Fig. 15. Performance under update (30 percent): ($uaree(\tilde{p})$ = 19.42 percent, $used(\tilde{p})$ = 0.24 percent.

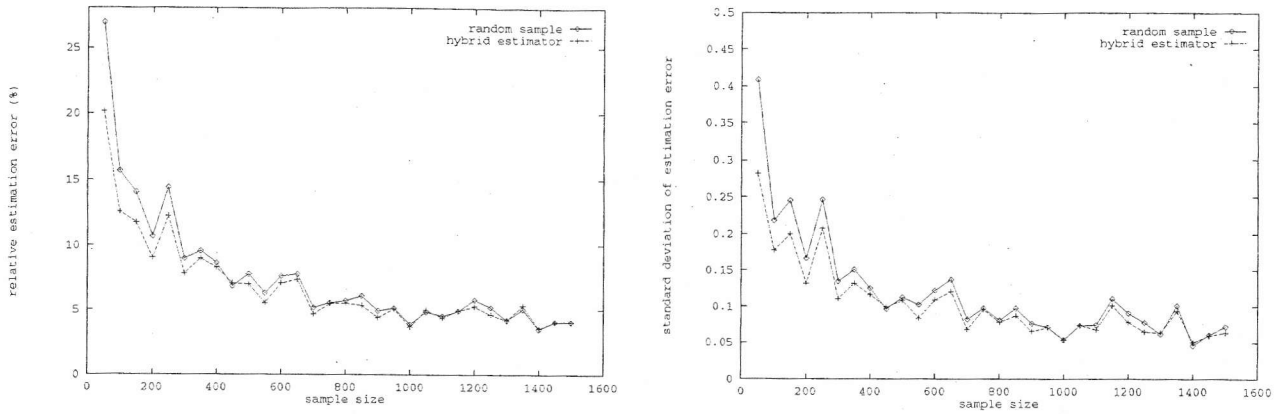

Fig. 16. Performance under update (50 percent): (*uaree*($\tilde{p}$) = 18.6 percent, *used*($\tilde{p}$) = 0.25 percent).

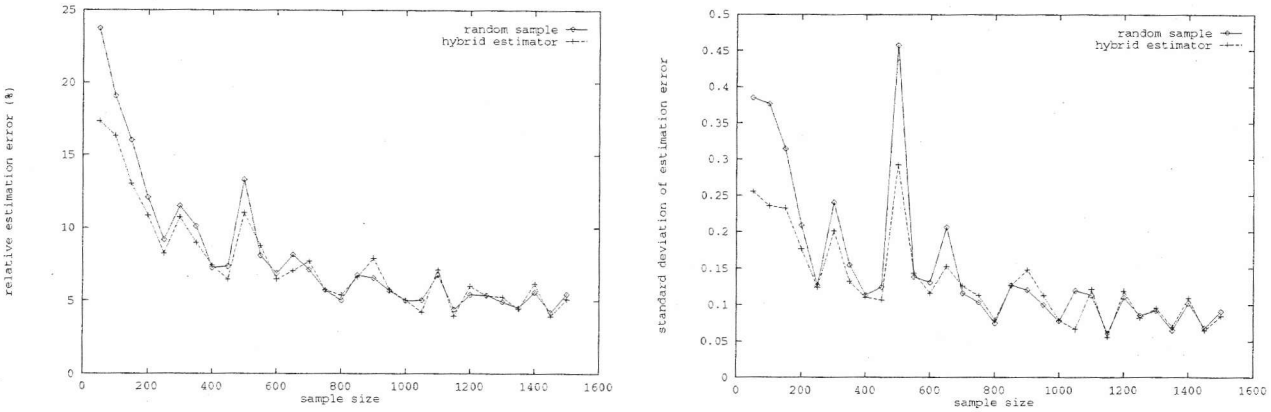

Fig. 17. Performance under insertion (5 percent): *uaree*($\tilde{p}$) = 21.38 percent, *used*($\tilde{p}$) = 0.30 percent).

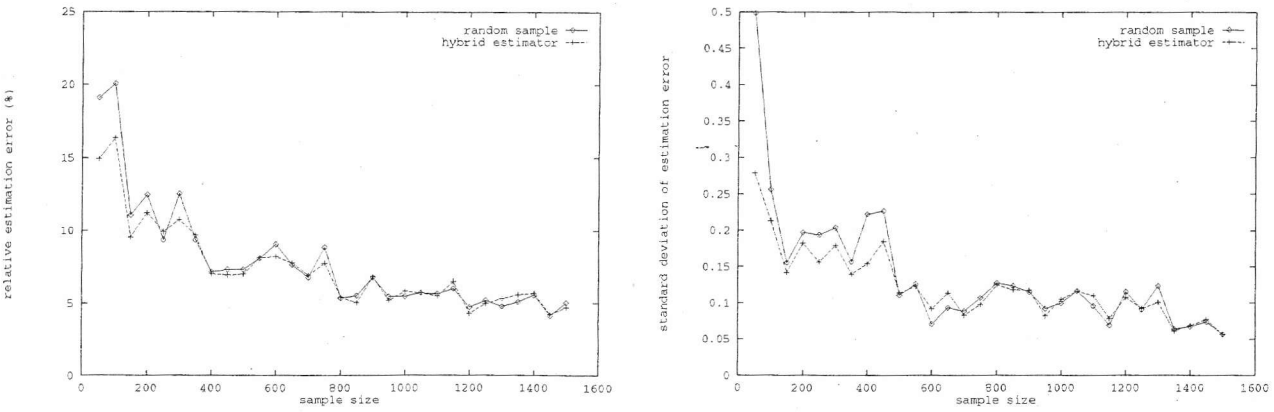

Fig. 18. Performance under insertion (10 percent): (*uaree*($\tilde{p}$) = 22.38 percent, *used*($\tilde{p}$) = 0.26 percent).

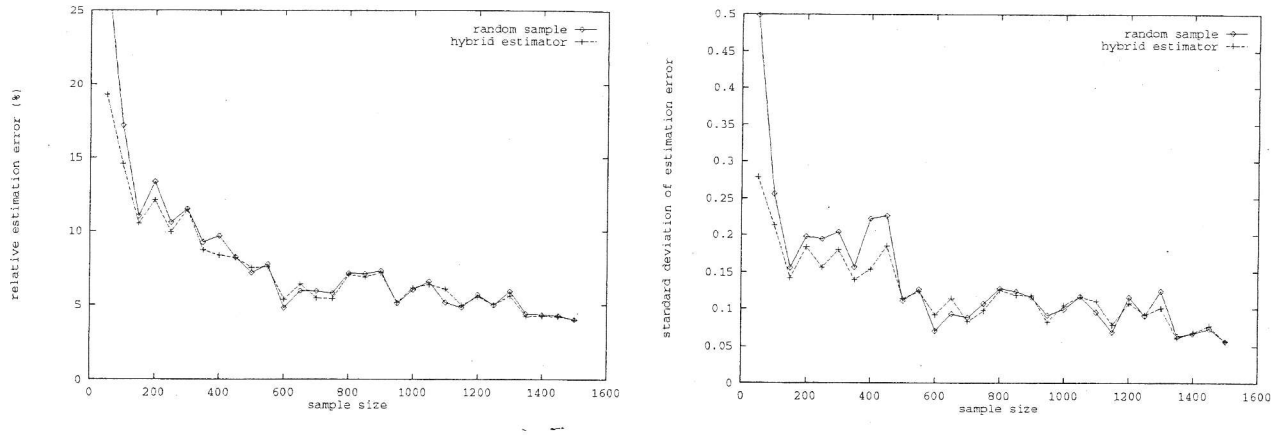

Fig. 19. Performance under insertion (20 percent): (*uaree*($\tilde{p}$) = 21.82 percent, *used*($\tilde{p}$) = 0.28 percent).

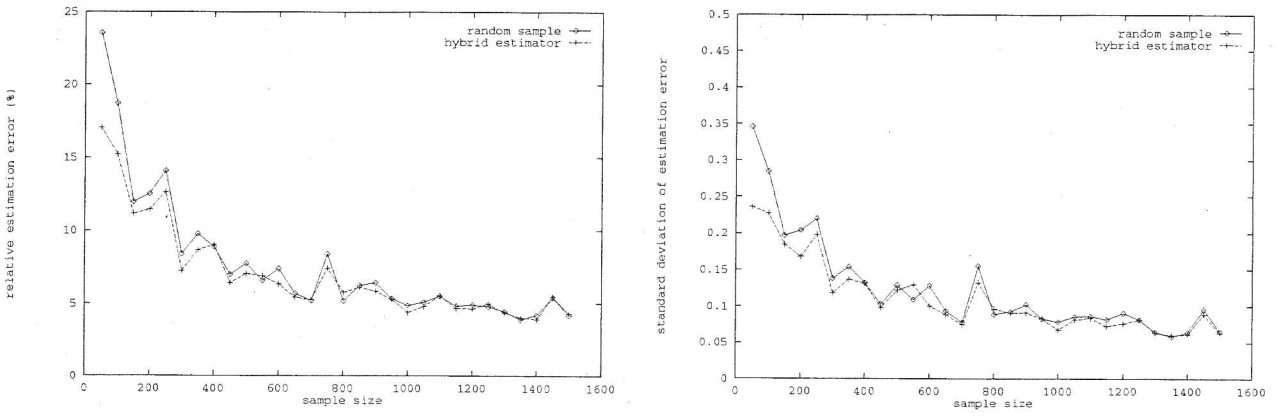

Fig. 20. Performance under insertion (50 percent): (*uaree*($\tilde{p}$) = 20.44 percent, *used*($\tilde{p}$) = 0.26 percent).

APPENDIX A

PROOF OF THEOREM 1:

By (1), the mean-squared error (*mse*) of $\breve{p}_n$ is written as:

$$E(\breve{p}_n - p)^2 = E((t \cdot (\hat{p}_n - p) + (1-t) \cdot (\tilde{p} - p))^2).$$

The above formula can be further simplified as:

$$E(\breve{p} - p)^2 = t^2 \cdot E(\hat{p}_n - p)^2 + (1-t)^2 \cdot (\tilde{p} - p)^2 = t^2 \cdot \frac{p \cdot (1-p)}{n} + (1-t)^2 \cdot (\tilde{p} - p)^2.$$

Let:

$$g(t) = t^2 \cdot E(\hat{p}_n - p)^2 + (1-t)^2 \cdot (\tilde{p} - p)^2 = t^2 \cdot \frac{p \cdot (1-p)}{n} + (1-t)^2 \cdot (\tilde{p} - p)^2. \quad (4)$$

Taking derivative of $g(t)$ with respect to t, we obtain:

$$g'(t) = 2 \cdot t \cdot \frac{p \cdot (1-p)}{n} - 2 \cdot (1-t)(\tilde{p} - p)^2. \quad (5)$$

Solving the equation $g'(t) = 0$, we get $t = t_n^*$. It follows that $E(\breve{p}_n - p)^2$ attains the minimum at t_n^*, which is of the form:

$$t_n^* = \frac{(\tilde{p} - p)^2}{p \cdot (1-p) / n + (\tilde{p} - p)^2}. \quad (6)$$

Substituting (6) into (4), we have:

$$E\left(\breve{p}_n^* - p\right)^2 = \frac{p \cdot (1-p)}{n} \cdot \frac{(\tilde{p} - p)^4 + p \cdot (1-p) \cdot (\tilde{p} - p)^2 / n}{\left[(\tilde{p} - p)^2 + p \cdot (1-p) / n\right]^2} = \frac{p \cdot (1-p)}{n} \cdot \left[1 - \frac{(\tilde{p} - p)^2}{p \cdot (1-p) / n + (\tilde{p} - p)^2}\right]. \quad (7)$$

That is, as long as $p \neq \tilde{p}$ and $0 < p < 1$, we have:

$$\frac{E\left(\breve{p}_n^* - p\right)^2}{E(\tilde{p} - p)^2} = \left(1 - t_n^*\right) < 1. \quad (8)$$

[16] P.J. Haas and A.N. Swami, "Sequential Sampling Procedures for Query Size Estimation," *Proc. SIGMOD*, pp. 341-350, ACM, 1992.
[17] P.J. Haas and A.N. Swami, "Sampling-Based Selectivity Estimation for Joins Using Augmented Frequent Value Statistics," *Proc. 11th Int'l Conf. Data Eng.*, pp. 522-531, Mar. 1995.
[18] W.-C. Hou and G. Ozsoyoglu, "Statistical Estimators for Aggregate Relational Algebra Queries," *ACM Trans. Database Systems*, vol. 16, no. 4, pp. 600-654, Dec. 1991.
[19] W.-C. Hou, G. Ozsoyoglu, and E. Dogdu, "Error-Constrained Count Query Evaluation in Relational Databases," *Proc. ACM SIGMOD Conf.*, pp. 278-287, Aug. 1991.
[20] Y.E. Ioannidis and S. Christodoulakis, "Optimal Histograms for Limiting Worse-Case Error Propagation in the Size of the Join Results," *ACM Trans. Database Systems*, vol. 18, no. 4, pp. 709-748, 1993.
[21] Y.E. Ioannidis and V. Poosala, "Balancing Histogram Optimality and Practicality for Query Result Size Estimation," *Proc. ACM SIGMOD Conf.*, pp. 233–244, 1995.
[22] Y. Ling, "Self-Organizing Model for Size Estimation in Database System," PhD dissertation, School of Computer Science, Florida International Univ., 1995.
[23] Y. Ling and B. He, "Entropic Analysis of the Biological Models," *IEEE Trans. Biomedical Eng.*, vol. 40, no. 12, pp. 1,193–1,200, Dec. 1993.
[24] Y. Ling and W. Sun, "A Supplement to Sampling-Based Methods for Query Size Estimation in a Database System," *SIGMOD Record*, vol. 16, pp. 15-18, Dec. 1992.
[25] Y. Ling and W. Sun, "A Size Estimator for Selections Based on Neural Network Learning Model," *Proc. Fifth Japanese Neural Network Soc. Conf.*, Tsukuba, Japan, Oct. 1994.
[26] Y. Ling and W. Sun, "A Comprehensive Evaluation of Sampling-Based Size Estimation Methods for Selections in Database Systems," *Proc. IEEE 11th Int'l Conf. Data Eng.*, pp. 532-539, Mar. 1995.
[27] Y. Ling, W. Sun, and X. Xiang, "A Hybrid Estimator: Full Utilization of All Available Information," technical report, Florida International Univ., 1995.
[28] R. Lipton and J. Naughton, "Estimating the Size of Generalized Transitive Closures," *Proc. 15th Int'l Conf. Very Large Data Bases*, pp. 165-172, 1989.
[29] R. Lipton and J. Naughton, "Query Size Estimation by Adaptive Sampling," *Proc. Ninth ACM Symp. Principles of Database Systems*, pp. 40-46, Mar. 1990.
[30] R. Lipton, J. Naughton, and D. Schneider, "Practical Selectivity Estimation Through Adaptive Sampling," *Proc. ACM SIGMOD*, pp. 1–11, 1990.
[31] M.-L. Lo, M.-S. Chen, C.V. Ravishankkar, and P. Yu, "On Optimal Processor Allocation to Support Pipelined Hash Joins," *Proc. SIGMOD*, pp. 69-78, 1993.
[32] C.A. Lynch, "Selectivity Estimation and Query Optimization in Large Databases with Highly Skewed Distributions of Column Values," *Proc. 14th VLDB Conf.*, pp. 240–251, 1988.
[33] M. Mannino, P. Chu, and T. Sager, "Statistical Profile Estimation in Database Systems," *ACM Computing Survey*, vol. 20, no. 3, pp. 191–221, Sept. 1988.
[34] J.F. Naughton and S. Seshadri, "On Estimating the Size of Projections," tech. report, Computer Dept., Univ. of Wisconsin–Madison, 1992.
[35] K.T. Sun and H.C. Fu, "A Hybrid Neural Network Model for Solving Optimization Problem," *IEEE Trans. Computer*, vol. 42, no. 2, pp. 218-227, Feb. 1993.
[36] W. Sun, Y. Ling, N. Rishe, and Y. Deng, "An Instant and Accurate Size Estimation Method for Joins and Selections in a Retrieval-Intensive Environment," *Proc. SIGMOD*, pp. 79–88, 1993.
[37] T. Wolf, D. Dias, and P. Yu, "An Effective Algorithm for Parallelizing Sort Merge Joins in the Presence of Data Skew," *Proc. Second Int'l Symp. Databases in Parallel and Distribution Systems*, pp. 103–115, 1990.
[38] J.M. Zurada, *Introduction to Artificial Neural Systems*. New York: West Publishing, 1992.

Yibei Ling received a PhD degree in computer science from Florida International University (FIU) in November 1995. He then worked as a research scientist at the High-Performance Database Center at FIU. In September 1997, he joined Dow Jones Markets (now Bridge Telerate) as a senior system designer, developing a Web-based trading system. In 1998, he joined the Bell Communications Research Laboratory (Bellcore) in Morristown, New Jersey. He has published several papers on query optimization in database systems, object-oriented methodologies, and biological models, including papers in *IEEE Transactions on Biomedical Engineering*, *IEEE Transactions on Knowledge and Data Engineering*, *Information Systems*, *ACM SIGMOD*, *Data Engineering*, and the *SIGMOD Record*. His primary research interests include query optimization in database systems, biological models, cost models, multi-tier system design, system architecture for Web-based trading systems, and real-time information systems.

Wei Sun received a PhD degree in computer science from the University of Illinois, Chicago, in August 1990. He then joined Florida International University (FIU), and is now an associate professor, director of the Multimedia Computing and Database System, and associate director of the High-Performance Database Research Center. Since 1989, he has co-authored five books and one book chapter; and he has published more than 50 technical papers, including papers in *ACM Transactions on Database Systems*, *IEEE Transactions on Knowledge and Data Engineering*, *Communications of the ACM*, *ACM SIGMOD*, and the *Computer Journal*. He served as general chair of the Fourth IEEE/ACM International Conference on Parallel and Distributed Information Systems (PDIS) in Miami Beach, Florida, in December 1996; and he has served on the program and organizational committees of more than a dozen international conferences. His major research interests are in database systems, Internet/Intranet technologies, knowledge-based systems, and multimedia systems. His research and development have been sponsored by NASA, the National Science Foundation, the United States Department of Agriculture, the Florida Department of Education, and private sectors. He is a senior member of the IEEE.

Naphtali D. Rishe received his PhD at Tel Aviv University in 1984. He was an assistant professor at the University of California, Santa Barbara (1984-1987); and he was a professor (1987-1992) and is now an associate professor (since 1992) at Florida International University (FIU). He is the founder and director of the High-Performance Database Research Center at FIU, chaired the Program and Steering Committees of the PARBASE Conference, and is on the Steering Committee of the PDIS Conference series. His work on the Semantic Binary Database Model was published as a book by Prentice Hall in 1988; and his semantic modeling theory was published as a book by McGraw Hill in 1992. His current research focuses on efficiency and flexibility of database systems (particularly object-oriented, semantic, decision-support, and spatial/geographic DBMS), distributed DBMS, high-performance systems, database design tools, and Internet access to databases. He has edited three books and authored 23 papers in journals (including *IEEE Transactions on Knowledge and Data Engineering*, *DKE*, *Information Systems*, and *Fundamenta Informaticae*), seven chapters in books and serials (including three in Springer-Verlag's Lecture Notes in Computer Science series), and more than 50 papers in conference and symposia proceedings (including ACM SIGMOD, PDIS, IEEE DE, ACM SIGIR, SEKE, ARITH, and FODO). He has been awarded millions of dollars in research grants by government and industry. His research is currently sponsored by NASA (4M), the National Science Foundation (3M), NATO, BMDO, ARO, DoD, DoI, and other agencies. He is a member of the IEEE Computer Society.

Xianjing Xiang received his PhD in statistics from the University of Chicago in 1992. He was an assistant professor at the University of Oregon (1992-1995), and he is currently a senior biostatistician at Novartis Pharmaceuticals in Summit, New Jersey. His research interests are statistics and statistical applications in database management. He is the author of 17 papers published in various journals, including the *Annals of Probability* and the *Annals of Statistics*.