Probability Distributions for Seek Time Evaluation

YANNIS MANOLOPOULOS

Division of Electronics and Computer Engineering, Department of Electrical Engineering, Aristotelian University of Thessaloniki, 54006 Thessaloniki, Greece

ABSTRACT

In this work, two known probability distributions based on the replacement model, namely, the Maxwell–Boltzmann and Bose–Einstein ones, are examined. Each of these distributions is assigned to a specific case of query processing, which is known in the literature as batching on primary key values or secondary key retrieval, respectively. These distributions are applied in evaluating the expected distance traveled by the read/write heads in disk searching. Two new probability distributions are derived, based on the previous ones, which are involved in the evaluation of the expected number of the cylinder hits. Therefore, seek time evaluation is faced globally. Numerical results are given.

1. INTRODUCTION

Query processing optimization is an integral part of recent database and knowledgebase systems and, mainly, concerns the minimization of I/O cost and the required space. For example, it arises as the final, at the physical level, phase for an efficient implementation of rule satisfaction subsystems [16] or the optimal ordering of conjunctive queries [17]. It should be noted that the queries may have various origins and degrees of complexity, e.g., queries based on single or range primary key values, secondary key retrievals, etc.

Query optimization may be viewed in two respects: the optimization of every query separately of the others as well as the optimization of a set of queries as a whole. At the physical level, the performance of any implementation depends on the maintenance algorithms and the structure of the specific files and indexes and, consequently, the number of requested records, the number of fetched pages, and the number of hit and traveled cylinders by the moving arm mechanism. This work reports on the use of some probability distributions involved in the performance evaluation of such systems under these cost metrics. In addition, these probability distributions are applied for the derivation of estimates on the number of cylinder hits and cylinders traveled. Thus, mathematical analysis, which is necessary for decision-making

655 Avenue of the Americas, New York, NY 10010 0020-0255/92/$03.50

on the techniques to be implemented, is supported. This kind of work has a direct application in query optimizers. Therefore, our results may be augment the list of rules, or may contribute to the derivation of new ones, of relational database query optimizers such as those reported in [14].

The rest of the paper is organized as follows. In Section 2 two specific query types are considered, namely, the ones that are known in the literature as batching on primary key values and secondary key retrieval. In addition, two replacement models are related to these query types. These replacement models are the Maxwell–Boltzmann and Bose–Einstein ones. Examples are given and the appropriate replacement model is assigned to each query type. In Section 3 the focus is on seek time evaluation when a disk with movable read/write heads is searched. These distributions are used to derive new formulas for the expected seek distance when the disk is searched under the SCAN scheduling policy. In this section, two new probability distributions, based on the previous ones, are introduced giving the expected cylinder visits as a function of the requests. Each one of these new probability distributions is related to the two specific query types. In Section 4 numerical results are illustrated. It is also discussed how probability distributions are related to the concept of reorganization. In that section the final conclusions are included.

2. QUERY TYPES AND REPLACEMENT DISTRIBUTIONS

Suppose we are given the relation:

EMP(empno, empname, age, sex, salary, deptno)

Assume also that a small number of requests of the following form arrives almost simultaneously:

"Retrieve the record of employee *X*"

where *X* is an employee code number. Another incoming set of user requests would be

"Retrieve the names of all the employees
aged under *Y* and salary less than *Z*"

where *Y* and *Z* are values belonging to the domain of these attributes.

It is worth noting that the chance that two users may pose identical or basically similar queries is not excluded. Here, the word similar insinuates that the queries differ only in the values of the parameters X, Y, Z. A direct consequence of this position is that replacement models should be used. This occurs irrespective of the cost metric used to charge the queries, e.g., the number of records required, the number of pages fetched, or finally the number of cylinders visited or hit.

Evidently, the difference between the two types of query is that the first (second) is based on primary (secondary) key values. However, both sets of queries may be answered as a whole instead of being processed on a first-come-first-served basis. This technique is advantageous from the user as well as the system point of view; e.g., the average response time is minimized and the computer resources are used more economically. The gain is due to the fact that excess accesses to the same physical pages are ommitted. Reference [10] notes all the previous works on this subject.

Here, it is noted that a common method in database systems to support the second query type is the use of secondary indexes. In the absence of these structures the set of primary key values is not known in advance and therefore there is no possibility for the determination of shared accesses. Secondary indexes provide every distinct attribute value with a list of the corresponding records in terms of their primary key values or their physical addresses [1]. There are certain fors and againsts in storing key values or addresses. Key values remain stable through the file lifetime; their addresses do not. In the second case maintenance, involving complexity, is necessary. On the other hand, addresses are usually kept in ascending order providing faster access. For answering the second query, type two secondary indexes must be visited and a merging of the two lists must be carried out to determine the actual records to be retrieved. If the address lists are not ascendingly ordered or if the key values are stored, then a sorting may be performed before merging. In the sequel we assume that the indexes contain addresses in ascending order.

In the following we concentrate on the comparison of a set of queries based on primary key values to a single query based on a secondary key value, since they may be answered in a similar way by a single scanning of the relation. The most important difference from the performance evaluation point of view is that the requested records of the first (second) case are (not) independent of each other. This phenomenon stems from the fact that some users may ask information about the same person X, while the records satisfying the secondary key based query are determined after accessing the secondary indexes which are ordered ascendingly. This fact may be better comprehended by the following example.

EXAMPLE. Let two balls, named A and B, fall into three urns, named 1, 2, and 3, respectively. Table 1, (2) shows the possible outcomes when the balls

TABLE 1

Balls Fall Independently

1	2	3
A B		
A	B	
A		B
B	A	
	A B	
	A	B
B		A
	B	A
		A B

(do not) fall independently. For example, note that in Table 2 there is only one sequence of balls: A followed by B but not B followed by A. This is due to the fact that secondary indexes give the addresses in ascending sequence. However, in Table 1 the case that ball A is thrown in urn 1 and ball B in urn 2 is different than the case that ball A is thrown in urn 2 and ball B in urn 1.

Now speaking more formally in computer terms, assume that a transaction requests n records from a file consisting of m pages, each with capacity N/m records, where N is the total number of records. A question arising is what is the expected number of pages containing the satisfying records, if the transaction is processed sequentially or randomly. Suppose, also, that the file occupies m cylinders. In this case, the question is what is the expected number of cylinders hits or what is the expected number of cylinders traveled since these quantities are critical in evaluating the seek time. It can be easily seen that the second problem is equivalent to the sequential case of the previous problem under the assumption that the process obeys the SCAN disk scheduling policy.

TABLE 2

Balls Do Not Fall Independently

1	2	3
A B		
A	B	
A		B
	A B	
	A	B
		A B

According to this policy, 'the requests are answered as the head travels alternatively from the outer to the inner cylinder of the disk.

It is very probable that there may exist duplicate records among the n requested ones. It is reasonable, also, to assume that the cylinders contain an infinite number of records. These are the reasons why a replacement model should be adopted to describe these processes. The literature on the two problems described above is rich. References [10–12] by Manolopoulos and Kollias contain pointers to earlier works on both the problems. More specifically, we note that the works of Cardenas [4], Cheung [5], Christodoulakis [6], Langer and Shum [9], and Manolopoulos and Kollias [10, 11] consider probability distributions based on the replacement model. What, to the author's knowledge, merely exists in the literature is a validation for the instances that a replacement model should be adopted as well as an interpretation of the distributions obeying this model and describing the two query types under examination.

There are two probability distributions which are based on the replacement model [13, p. 13]. For example, according to the Maxwell–Boltzmann model (MB model) the number of possible outcomes is m^n, but according the Bose–Einstein model (BE model) the number of possible outcomes is $C(m+n-1,n)$, where $C(a,b)$ is the number of the a-choose-b combinations. Table 3 gives the number of possible outcomes which are produced under the assumptions that either distinguishable or not distinguishable particles are selected, either with or without replacement. Selection without replacement is depicted for the purpose of completeness but is not met in the environment under consideration.

The Maxwell–Boltzmann (Bose–Einstein) model corresponds to the example of Table 1 (2). Therefore, the MB and BE distributions are valid for specific query cases, namely, the two query types described in the beginning of this section. The reason is reminded again: the MB (BE) model assumes that the particles are (not) distinguishable. This fact should be interpreted as if the

TABLE 3

Number of Possible Outcomes and Corresponding Model When Particles Distinguishable or Not Are Selected With or Without Replacement

	Distinguishable particles	Nondistinguishable particles
With replacement	m^n outcomes Maxwell–Boltzmann	$C(m+n-1,n)$ outcomes Bose–Einstein
Without replacement	no physical reality	$C(m,n)$ outcomes Fermi–Dirac

records are (not) independent. We proceed now by applying these distributions in the context of disk searching.

3. SEEK TIME EVALUATION

It is known that when a disk is searched, the seek cost is the dominating cost factor when compared to the rotational delay and the transfer time. For example, the seek time for moving from one cylinder to another is given by the following equation [8]:

$$\mathrm{Tsk} = S_{\min} + s * d, \tag{1}$$

where $S_{\min}$ is the time required to move to the next cylinder, d is the distance between the two cylinders, and s is a constant and equals the quotient of the difference $(S_{\max} - S_{\min})$ by $(m-1)$, where m is the number of cylinders and $S_{\max}$ is the time required to move from the first to the last cylinder. A physical explanation of the quantity $S_{\min}$ is due to the inertia of the disk arm; in other words, it gives the startup time. Suppose that the n requests are to be satisfied and that the disk searching obeys the SCAN scheduling policy, according to which the disk heads travel alternatively from the inner to the outer cylinder. Therefore, the total seek time is given by the summation [8]

$$\begin{aligned}\sum_{i=1}^{n} \mathrm{Tsk}_i &= \sum_{i=1}^{n} [S_{\min} + s * d_i] \\ &= v * S_{\min} + s * \sum_{i=1}^{n} d_i, \end{aligned} \tag{2}$$

where v gives the number of cylinder hits by the n requests.

First, let us focus on the second quantity of formula (2). The evaluation of this quantity is based on the probability distribution of the lengths of the distances d_i's as a function of the number of the traveled cylinders and the total number of cylinders. If the first (second) query type is considered, then the following probability distribution is used [5, 6]:

$$P_1(k) = (k^n - (k-1)^n)/m^n, \tag{3a}$$

$$\begin{aligned}P_2(k) &= (C(n+k-1,n) - C(n+k-2,n))/C(n+m-1,n) \\ &= C(n+k-2,n-1)/C(n+m-1,n)\end{aligned} \tag{3b}$$

These relations give the probability that all the n records satisfying the requests are stored in the first k cylinders exactly. This comes from the fact that the first and second terms of the numerator give the number of ways that the n requested records may be selected from the first k and $(k-1)$ cylinders, respectively. Therefore, the probability that the n requests are selected from the first $(k-1)$ cylinders and not the k cylinders is excluded. As in [3, 8, 11, 12], the following assumptions are the basis for the forthcoming two theorems and one corollary,

a. The relation occupies m consecutive cylinders.

b. The query processing program is dedicated to the satisfaction of each of the n queries, e.g., the possibility of changing the direction that the heads move to serve some other system request is excluded.

c. The disk head initially is positioned on top of the first (outer) cylinder occupied by the relation and is ready to move towards the last (inner) cylinder.

THEOREM 1. *The expected distance, in cylinders, traveled by the disk heads for the two query types respectively is*

$$\sum_{k=2}^{m} (k-1)P_1(k) = m - \frac{1}{m^n} \sum_{r=1}^{m} r^n \tag{4a}$$

$$\sum_{k=2}^{m} (k-1)P_2(k) = (m-1)n/(n+1) \tag{4b}$$

Proof. The proof of formula (4b) exists in [11]. The proof of formula (4a) follows. It is accepted that the read/write head is initially positioned over the first cylinder. Therefore, the expected distance is given by the left hand expression of formula (4a). By replacing the probability function with formula (3a) and simplifying the resulting relation, it is derived that

$$\begin{aligned}\sum_{k=2}^{m} (k-1)P_1(k) &= \sum_{k=2}^{m} (k-1)\left(k^n-(k-1)^n\right)/m^n \\ &= 1(2^n-1^n)/m^n + 2(3^n-2^n)/m^n + 3(4^n-3^n)/m^n \\ &\quad + \cdots + (m-1)\left(m^n-(m-1)^n\right)/m^n \\ &= (m-1) - \frac{1}{m^n}\sum_{r=1}^{m-1} r^n = m - \frac{1}{m^n}\sum_{r=1}^{m} r^n \qquad \blacksquare\end{aligned}$$

COROLLARY. *Formula (4a) may be approximated by*

$$mn/(n+1) - 1/2. \tag{5}$$

Proof. This follows easily by accepting that for large values of m [2]:

$$\frac{1}{m^n}\sum_{r=1}^{m} r^n = \frac{m}{(n+1)} + \frac{1}{2}. \qquad \blacksquare$$

However, the product $v * S_{\min}$ contributes substantially to the final result and needs to be calculated. We report now a second theorem providing the probability distribution functions for the calculation of v for the two query types.

THEOREM 2. *The probability distributions of the number of disk cylinder hits by requests of the two query types respectively are*

$$P_3(k) = \left[\left(\frac{k}{m}\right)^n - \sum_{i=1}^{k-1} (-1)^{i+1}\left(\frac{k-i}{m}\right)^n C(k, k-i)\right] C(m,k) \tag{6a}$$

$$P_4(k) = \left[C(k+n-1, n) - \sum_{i=1}^{k-1} (-1)^{i+1} C(k-i+n-1, n) C(k, k-i)\right]$$
$$\times \frac{C(m,k)}{C(m+n-1,n)}, \tag{6b}$$

where m is the number of cylinders, n is the number of records satisfying the request, and k is the number of cylinder hits $(1 \leqslant k \leqslant n)$.

Proof. Part 1 for formula (6a): Following formula (3a) the probability that the n requests will follow in k cylinders is $(k/m)^n$. This probability contains also the probability that the k requests will follow in $(k-1)$ cylinders or even less. By using inclusion and exclusion the summation on i is derived. This difference is multiplied by the possible number of selections that the k cylinders may be chosen out of the m cylinders of the disk.

Part 2 for formula (6b): The combination $C(m+n-1, n)$ in the denominator gives the number of possible outcomes when the n records are selected with replacement from m cylinders. The combination $C(m,k)$ in the numerator is the possible number of selections of the k cylinders out of the m cylinders of the disk. The quantity $C(k+n-1, n)$ gives the possible number of combinations when the n records may fall with replacement in these k cylinders. The latter quantity does not quarantee that the n records will fall in exactly k cylinders. The derivation of the summation comes easily by consider-

TABLE 4

Expected Distances Traveled by Using Formulas Based on BE Model, MB Model, and Approximation of MB Model

m, n	Formula (4b)	Formula (4a)	Formula (5)
100, 5	82.50	82.83	82.83
100, 10	90.00	89.40	90.41
100, 15	92.81	3.24	3.25
400, 5	332.50	332.83	332.83
400, 10	362.72	363.13	363.13
400, 15	374.06	374.50	374.50

ing an inclusion–exclusion technique. Therefore, by subtracting the summation it is guaranteed that the n records will fall in exactly k cylinders. Relation (6b) follows. ■

For both cases the expected number of cylinder hits, that is v, is equal to

$$v = \sum_{k=1}^{n} kP(k) \tag{7}$$

where $P_3(k)$ or $P_4(k)$ should replace $P(k)$. Formula (7) has not been possible to be simplified.

4. DISCUSSION AND CONCLUSION

To show practically the difference of the two pairs of probability distributions, we give the following tables with numerical results. Table 4 {5} demonstrates the results of relations (4a, 4b, 5) {(6a, 6b, 7)} under varying parameters m and n.

TABLE 5

Expected Cylinder Hits by Using Formulas Based on MB Model and BE Model

m, n	Formulas (6a)–(7)	Formulas (6b)–(7)
100, 5	4.90	4.81
100, 10	9.56	9.17
100, 15	13.99	13.16
400, 5	4.98	4.95
400, 10	9.89	9.78
400, 15	14.74	14.49

In Table 4 it is remarked that formula (5) is a very close approximation of the exact one (4b). Therefore, it can be used in place of the exact relation since it is computationally inexpensive. Also, results derived by using the BE model are always smaller but very close to those derived by using the BM model. In Table 5 we make the same remark as in Table 4, i.e., the results derived by using the BE model are always smaller than these derived by using the MB model. Also, for a constant number of requests (cylinders) the absolute and the relative difference between the two models grows with decreasing (increasing) number of cylinders (requests). Simulation results are close to the analytic ones depicted in both tables and, therefore, are omitted.

It is known that "the entropy of a probability distribution is larger as the distribution is nearer to uniform and as the possible number of values grows larger " [7]. It is known, also, "that the entropy of probability distributions is a Schur concave function" and therefore "if a probability distribution p majorizes another distribution p', then p' has greater entropy than p." It can be shown by some simple experiments on the two pairs of formulas (3a)–(3b) and (6a)–(6b) that the distributions based on the BE model are more uniform than the ones based on the MB model. Therefore, it is anticipated that the observation on MB and BE models may be formally proved by using majorization theory along the lines of the last reference.

We return again to discuss the issue of secondary indexes. Suppose that the secondary indexes are created at the same time with the creation of the file. In addition, suppose that for every distinct attribute value the secondary index contains a list with the relevant primary key values (addresses of the corresponding records) ordered in ascending order of these values (addresses). Through the file lifetime insertions, deletions, and updates are performed. Therefore, the indexes are updated and after a time interval the lists become unordered [15]. The penalty of either sorting the addresses or accessing the records by using their primary key values has to be paid before accessing the requested records. In conjunction with Tables 1 and 2, we remark that this price has to be paid because the distribution of the addresses on the domain of the cylinder values starts with a form similar to Table 2 and degenerates to a form similar to Table 1. Reorganization is performed periodically in order to restore the physical contiguity of the file and the corresponding indexes and, as a consequence, to eliminate the effects of their updates on the disordering of the addresses.

In the past a lot of work has appeared in the literature for the optimization of database query processing and the corresponding performance evaluation at physical level, e.g., the estimation of the number of required records by a transaction, the number of page accesses in secondary storage, as well as the number of cylinders traveled by the read/write heads when the disk is searched by using the SCAN algorithm. Replacement and non replacement models have been used.

In this study we explicitly specified under what conditions the two models of replacement selections (Maxwell-Boltzmann and Bose-Einstein) may be used for evaluating the cost of a query. More precisely, the MB model or BE model should be used for what is called in the literature batching on primary key values or secondary key retrieval, respectively. In the first (second) case requested records (do not) arrive independently. These distributions are applied for evaluating the expected seek distance traveled by the disk heads. In addition, a computationally inexpensive simplifying formula is given to be used in the place of the one based on the MB model.

Crucial, also, in the evaluation of the total seek time is the number of cylinder visits for the two types of queries examined. In this work, two new probability distributions, which are based on the MB and BE models, are provided. These distributions give the number of cylinder hits as a function of the number of the cylinders and the required records. New derived formulas give the expected number of cylinder hits. Numerical results are given showing that the difference under the two models is not great.

Thanks are due to the anonymous referee for his remarks on the first version of this work.

REFERENCES

1. H. D. Anderson and P. B. Berra, Minimum cost selection of secondary indexes and formatted files, *ACM Trans. Database Syst.* 2(1):68–90 (1977).
2. W. H. Beyer, *Standard Mathematical Tables*, 25th ed., CRC Press, Boca Raton, Fl., 1979.
3. F. W. Burton and J. G. Kollias, Optimising disk head movements in secondary key retrievals, *Comput. J.* 22(3):206–208 (1979).
4. A. F. Cardenas, Analysis and performance of inverted database structures, *Commun. ACM* 18(5):253-263 (1975).
5. T. Y. Cheung, Estimating block accesses and number of records in file management, *Commun. ACM* 25(7):484-487 (1982).
6. S. Christodoulakis, Estimating Block Transfers and Join Sizes, *Proceedings of ACM SIGMOD-83 Conference*, pp. 40–54, 1983.
7. S. Christodoulakis, Implications of certain assumptions in database performance evaluation, *ACM Trans. Database Syst.* 9(2):163–186 (1984).
8. J. G. Kollias, An estimate of seek time for batched searching of random and index sequential structured files, *Comput. J.* 21(2):132–133 (1978).
9. A. M. Langer and A. W. Shum, The distribution of granule accesses made by database transactions, *Commun. ACM* 25(11):831–832 (1982).
10. Y. Manolopoulos and J. G. Kollias, Expressions for partly and completely unsuccessful batched search in sequential and tree structure files, *IEEE Trans. Software Eng.* 15(6):794–799 (1989).

11. Y. Manolopoulos and J. G. Kollias, Disk head movement in batched searching, BIT 28:27–36 (1988).
12. Y. Manolopoulos and J. G. Kollias, Performance of a two-headed disk system when serving database queries under the SCAN policy, *ACM Trans. Database Syst.* 14(3):425–442 (1989).
13. A. Papoulis, *Probability, Random Variables and Stochastic Processes*, McGraw-Hill, New York, 1965.
14. B. Saltzberg, *File Structures*, Prentice-Hall, Englewood Cliffs, N.J., 1988.
15. G. H. Sockut and R. P. Goldberg, Database reorganization—principles and practice, *ACM Comput. Surv.* 11(4):371–395 (1979).
16. J. D. Ullman, *Principles of Database and Knowledgebase Systems*, Computer Science Press, 1988.
17. C. K. Wong, Minimizing expected head movement in one-dimensional and two-dimensional mass storage systems, *ACM Comput. Surv.* 12(2):167–178 (1980).

Received 13 July 1988; revised 30 May 1989