

ОПТИМИЗАЦИЯ ФИЗИЧЕСКИХ СТРУКТУР БАЗ ДАННЫХ АСУ

Р. ЖИЛИНСКАС

В условиях применения современной вычислительной техники для решения задач АСУ возникает необходимость полного и рационального использования средств обработки информации. Абстрагируясь от конкретного содержания данных, можно выделить некоторые параметры, характеризующие АСУ. Они определяют технологию обработки данных и базу данных, хранимую в памяти ЭВМ.

Эффективность функционирования баз данных (БД) определяется степенью удовлетворения потребностей пользователей и затратами на организацию и ведение БД. В первом случае важным элементом эффективности БД является время реакции БД на запрос (время ответа). Это время во многом зависит от физической структуры БД. При выборе оптимальной физической структуры БД необходимо учитывать эффективное использование памяти ЭВМ; время доступа к данным; время выборки данных; удаление избыточности; рациональное внесение изменений.

Наиболее подходящим методом для оптимизации структуры БД является статистико-оптимизационный, так как получить достоверные динамические характеристики БД можно лишь собрав статистические сведения о динамике данных базы. Проблемы ведения статистики БД и выбор решающих оптимизационных параметров представлены в работе (2). Ниже исследуются проблемы оптимизации физических структур БД, касающиеся минимизации времени доступа к данным и времени выборки данных.

Определим основные параметры, позволяющие установить физические характеристики файлов БД. Основными и наиболее распространенными параметрами являются объем файла (L), длина записи файла (Φ), коэффициент блокирования (β), коэффициент использования записей (α), коэффициент использования блоков (γ), коэффициент использования носителя информации (ρ). Первые два параметра легко определяются при формировании каждого конкретного файла. Следует лишь отметить, что L измеряется количеством записей в файле, а Φ — байтами. Коэффициент блокирования понимается как число, определяющее количество записей в блоке. От значения этого коэффициента во многом зависит время доступа и выборки данных и эффективное использование памяти ЭВМ. Параметры L , Φ , β выражают статическую характеристику БД.

Параметры α , γ , ρ выражают динамическую характеристику БД, так как их значения зависят от конкретного случая использования данных. Коэффициент использования записей (α) в общем случае выражает интенсивность использования данных. В некоторых работах (3; 4) он определяется как соотношение среднего количества записей, использованных за период времени, и объема файла. Однако при использовании БД нас интересует время ответа, поэтому коэффициент α целесообразно опре-

делить для каждого конкретного случая обращения к БД. Коэффициент использования записей следует определить как отношение количества используемых записей за время обработки одного запроса к объему файла. Необходимо отметить, что разные запросы используют разное количество элементов записи одного и того же файла. Поэтому для БД введем понятие коэффициента использования одной записи (δ), определяющего отношение количества используемых знаков записи некоторым запросом к общему количеству знаков в записи.

При обращении к файлам БД в оперативную память (ОП) ЭВМ вводятся блоки файлов. Разное количество используемых записей в разных запросах влечет использование разного количества блоков. Для определения этого количества применяется коэффициент использования блоков (γ). Так как заранее практически нельзя установить, в каких блоках находятся искомые записи, будем считать, что они распределены по файлу случайно, а найти любую запись можно с одинаковой вероятностью. Рассматривая файл как некоторую прямую, точки на которой определяют позиции искомым записей, можно определить вероятность использования тех или иных отрезков (блоков) прямой (файла), считая, что вероятность обращения к блоку не зависит от вероятности обращения к предыдущему блоку.

Пусть файл разбит на блоки величиной по β записей. Коэффициент использования записей равен α . Тогда вероятность того, что в блоке (отрезке прямой) не будет ни одной из искомым записей, равна $e^{-\alpha\beta}$ (распределение Пуассона) (5). Вероятность того, что в блоке будет хотя бы одна искомая запись, равна

$$\gamma = 1 - e^{-\alpha\beta}.$$

Значение этой вероятности определяет значение коэффициента использования блоков.

Эффективность использования внешней памяти ЭВМ определяется с помощью коэффициента использования носителя информации. Имея в виду то обстоятельство, что основным носителем информации файлов БД является магнитный диск (МД), определим коэффициент использования цилиндров МД (ρ). Руководствуясь аналогичным рассуждением, как и при определении коэффициента использования блоков, устанавливаем, что коэффициент использования цилиндров равен

$$\rho = 1 - e^{-\alpha C},$$

где C — объем цилиндра в записях.

Оценка оптимальности структур файлов БД связана с определением времени и стоимости обработки файлов. Обработку файла составляют поиск начала файла, поиск цилиндра, поиск блока, обмен данными между внешней памятью и ОП, обработка данных процессором. Все эти составляющие обработки файла можно оценить по времени и стоимости обработки. Если принять, что файлы заранее определены по носителям информации, а затраты на эксплуатацию устройств внешней памяти зависят от времени обработки файлов, то критерием оптимальности физической структуры файлов БД можно выбрать время обработки файлов. Стоимость их обработки в общем случае можно выразить как произведение времени обработки и стоимости кванта времени процессора. При выборе методов организации файлов должна учитываться величина поверхности носителя информации, требуемая для записи файла, которая оценивается при сравнении с минимальной поверхностью, необходимой для данного типа файла.

Время t обработки файлов можно выразить следующим образом:

$$t = t^H + t^D + t^B + t^O + t^T,$$

где t^H , t^D , t^B , t^O , t^P — соответственно время поиска начала файла, время поиска цилиндров, время поиска блоков на цилиндре, время обмена данными между внешней памятью и ОП, время обработки данных процессором.

Время t является многообразным, зависящим от множества факторов. На рисунке представлена структурная схема зависимости времени t от физических характеристик файлов БД.

Как видно из рисунка, время поиска начала файлов зависит от организации потока запросов и размещения файлов на носителе информации. Так как в общем случае порядок поступления запросов является случайным, можно считать, что время t зависит от положения считывающей головки накопителя на МД по отношению к началу носителя. Тогда для каждого запроса можно принять, что время t равно времени подвода считывающей головки от начала носителя к его середине.

Время поиска цилиндров зависит от структуры файла и метода поиска данных, от размещения файлов на носителях информации, от количества используемых цилиндров, времени перемещения считывающей головки накопителя на МД. Время перемещения считывающей головки зависит от среднего времени одного перемещения считывающей головки и общего количества ее перемещений. Так как БД АСУ требуют единично произвольной и групповой произвольной выборки данных, большинство современных БД организуются с помощью индексно-последовательного способа записи данных на носитель информации. Поэтому при поиске данных в файле считывающая головка перемещается от индексной таблицы к основной области и обратно, а также между цилиндрами основной области. Количество перемещений считывающей головки от индексной таблицы к основной области и обратно равно

$$K^1 = 2 \frac{L^T}{\beta^T} \left[1 - \left(1 - \frac{\beta^T \gamma^O}{L^T} \right)^{\gamma^O \frac{L^O}{\beta^O}} \right] - 1,$$

где L^T — объем индексной таблицы;

β^T — коэффициент блокирования индексной таблицы;

γ^O — коэффициент использования блоков основной области;

L^O — объем основной области;

β^O — коэффициент блокирования основной области.

Количество перемещений считывающей головки между цилиндрами основной области составляет

$$K^2 = \rho^O \frac{L^O}{C^O} (1 + P) - \gamma^O \frac{L^O}{\beta^O},$$

где ρ^O — коэффициент использования цилиндров основной области;

C^O — количество записей на цилиндре основной области;

P — вероятность того, что информация об одном цилиндре находится больше, чем в одном блоке индексной таблицы.

Среднее время одного перемещения считывающей головки зависит от средней дистанции между индексной таблицей и основной областью (6) и определяется с помощью технических характеристик конкретного накопителя на МД. Тогда время поиска дорожек равно

$$t^D = \sum_{j=1}^m f_j \sum_{i=1}^n a_{ij} (K_{ij}^1 t_{ij}^1 + K_{ij}^2 t_{ij}^2),$$

где f_j — частота поступления запроса j за период времени;

$a_{ij} = \begin{cases} 1, & \text{если файл } i \text{ используется запросом } j, \\ 0 & \text{в противном случае,} \end{cases}$

t_{ij}^1 — среднее время одного перемещения считывающей головки от индексной таблицы к основной области и обратно при обра-

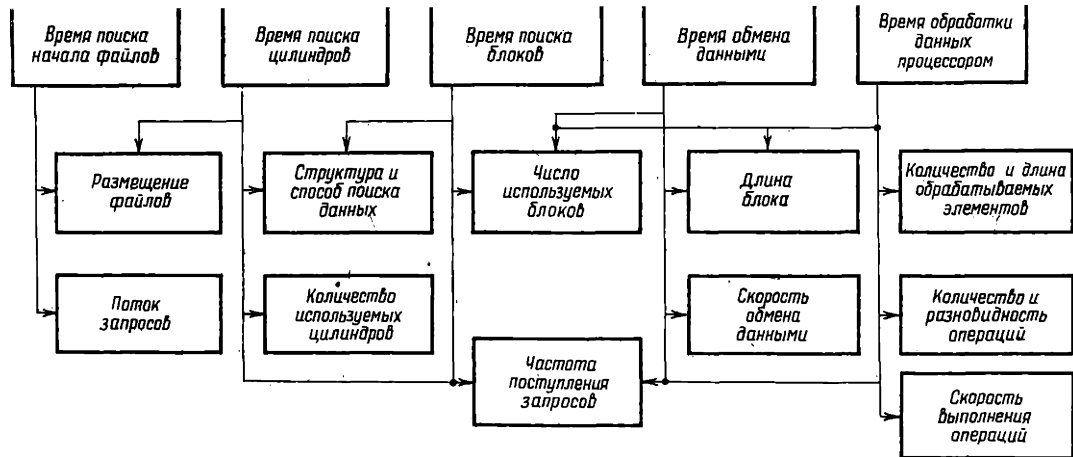


Рис. Структурная схема зависимости времени обработки файлов

ботке файла i запросом j ;

t_{ij}^2 — среднее время одного перемещения считывающей головки между цилиндрами основной области при обработке записей файла i запросом j ;

m — количество разновидностей запросов;

n — количество файлов в БД.

Время поиска блоков на дорожке зависит от количества используемых блоков, определяемых выражением $\gamma_i \frac{L_i}{\beta_i}$ (i — код файла), от частоты использования файла, ротационного времени t_i^B , определяемого по техническим характеристикам накопителя на МД. Время поиска блоков в общем случае можно определить следующим образом:

$$t^B = \sum_{j=1}^m f_j \sum_{i=1}^n \gamma_{ij} \frac{L_i}{\beta_i} t_i^B$$

где γ_{ij} — коэффициент использования блоков файла i запросом j .

Исходя из того обстоятельства, что величину блоков ограничивает объем ОП, предназначенный для ввода—вывода блоков, вступает в силу следующее ограничение:

$$\sum_{i=1}^n \alpha_{ij} \Phi_{ij} \beta_i \leq P_j, \quad (1)$$

где P_j — объем ОП, предназначенный для ввода—вывода данных при запросе j ;

Φ_i — длина записи файла i .

Так как величина блока не должна превышать объема дорожки, вступает в силу следующее ограничение:

$$\Phi_i \beta_i \leq C_i, \quad (2)$$

где C_i — объем дорожки (в байтах) МД, на котором расположен файл i .

Время обмена данными между внешней памятью и ОП зависит от количества вводимых в ОП блоков, количества, длины блока, частоты использования файла, скорости обмена данными между внешней памятью и ОП. Следовательно,

$$t^O = \sum_{j=1}^m f_j \sum_{i=1}^n \gamma_{ij} \frac{L_i}{\beta_i V_i^O} l_i,$$

где V_i^O — скорость обмена данными файла i между внешней памятью и ОП;

l_i — длина блока файла i (в байтах).

Время обработки данных процессором зависит от количества вводимых в ОП блоков, количества и длины элементов записей, участвующих в обработке запроса, от количества и разновидностей выполняемых операций, скорости выполнения операций. Количество и длину элементов записей, участвующих в обработке, можно определить с помощью коэффициента использования одной записи δ и коэффициента использования одного блока η_i . В работе (1) с помощью статистических испытаний доказано, что

$$\eta_i = \begin{cases} \alpha_i / 0,7, & \text{если } \alpha_i \leq 0,7, \\ 1 & \text{в противном случае.} \end{cases}$$

Количество и разновидность операций можно установить из алгоритма обработки запроса. В общем случае время обработки данных файлов процессором определяется следующим образом:

$$t^{\Pi} = \sum_{j=1}^m f_j \sum_{i=1}^n v_{ij} \frac{L_i}{\beta_i V_{ij}^{\Pi}} l_i \delta_{ij} \eta_{ij} d_{ij},$$

где δ_{ij} — коэффициент использования одной записи файла i запросом j ;
 η_{ij} — коэффициент использования одного блока файла i запросом j ;
 d_{ij} — количество операций над элементами записи файла i при обработке запроса j ;
 V_{ij}^{Π} — средняя скорость выполнения операций над элементами файла i при обработке запроса j .

Функция цели оптимизации физической структуры баз данных принимает следующий вид:

$$t = t^H + t^D + t^B + t^O + t^{\Pi} \rightarrow \min \quad (3)$$

при ограничениях (1), (2).

Как видно из выражений элементов функции цели (3), основными ее переменными являются β_i и C_i . Так как $C_i = \lfloor \frac{A_i}{\Phi \beta_i} \rfloor$, где A — ближайшее целое, не превышающее A , а ограничения накладываются лишь на величину β_i , за оптимизационную переменную принимается β_i . Ввиду огромной комбинаторности данной задачи оптимальное значение t в явном виде получить довольно сложно. Поэтому целесообразно минимизировать отдельные слагаемые выражения (3), используя метод градиентов. После этого полученные для каждого слагаемого оптимальные значения β_i сопоставляются и для каждого файла выбираются максимальные значения, удовлетворяющие ограничениям (1), (2). Таким образом получается практически оптимальное решение задачи оптимизации времени доступа к данным и выборки данных, имеющей функцию цели (3) и ограничения (1) и (2).

Вильнюсский университет
им. В. Капсукаса
Кафедра экономической
информации

Редколлегия вручено
в октябре 1982 г.

ЛИТЕРАТУРА

1. Жилинскас Р. Г. Сегментация файлов, обрабатываемых на малых ЭВМ.— Приборы и системы управления, 1979, № 10.
2. Жилинскас Р. Г. Использование статистического файла для ведения динамического информационного фонда АСУ.— В кн.: Вычислительные системы и АСУ. Вып. 3.— Вильнюс, 1982.
3. Мядоу Ч. Анализ информационных систем.— М.: Прогресс, 1977.
4. Померанцева И. В., Крюков Б. В. Исследование факторов, влияющих на организацию массивов.— В кн.: Проектирование технологии и организация обработки данных в АСУ. М.: МЭСИ, 1975.
5. Феллер В. Введение в теорию вероятностей и ее приложения. Т. 2.— М.: Мир, 1967.
6. Waters S. J. Estimating magnetic disc seeks.— The Computer Journal, 1975, vol. 18, No 1.